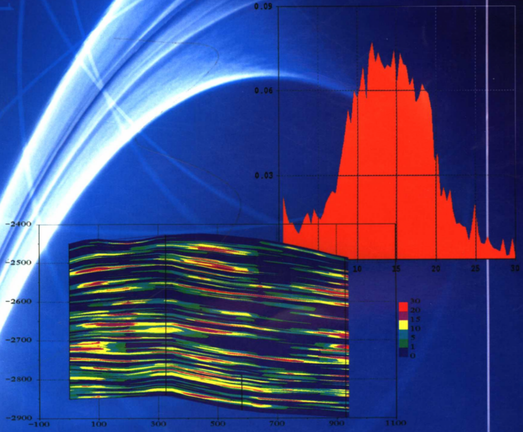
高等学校教材

现代数学地质

Modern Mathematical Geology

康永尚 沈金松 谌卓恒 编著



石油工業出版社

现代数学地质

现代数学地质

责任编辑：徐秀澎
封面设计：杨　群

ISBN 7-5021-5097-8
9 787502 150976 >

ISBN 7-5021-5097-8 / TE · 3944(课)
定价：16.80元

P628
3

高等学校教材

现代数学地质

康永尚　沈金松　谌卓恒　编著

石油工业出版社

内 容 提 要

本教材是针对油气勘探开发人才的培养而设计的，内容覆盖面广，从概率论和数理统计的基础知识到其应用，从单变量分析到多变量分析，从空间无关的变量到空间相关的变量，从线性科学的思维观到非线性科学的思维观，从易到难渐次深入，构成一个油气地质和开发领域从基础到应用的数学地质方法体系。

本书可作为高等院校地质勘探、地质工程及环境工程相关专业本科生和研究生的专业基础课教材，也可供石油地质等专业的生产和科研单位相关人员参考。

图书在版编目（CIP）数据

现代数学地质/康永尚等编著.
北京：石油工业出版社，2005.6
高等学校教材
ISBN 7-5021-5097-8

Ⅰ.现…
Ⅱ.康…
Ⅲ.数学地质-高等学校-教材
Ⅳ.P628

中国版本图书馆 CIP 数据核字（2005）第 051970 号

现代数学地质
康永尚　沈金松　谌卓恒　编著

出版发行：石油工业出版社
（北京安定门外安华里 2 区 1 号　100011）
网　址：www.petropub.com.cn
总　机：(010)64262233　发行部：(010)64210392
经　销：全国新华书店
印　刷：北京华正印刷厂

2005 年 6 月第 1 版　2005 年 6 月第 1 次印刷
开本：787×1092 毫米　开本：1/16　印张：11.75
字数：290 千字　印数：1—2000 册

定价：16.80 元
（如出现印装质量问题，我社发行部负责调换）
版权所有，翻印必究

前　　言

地质学最重要的应用是各种自然资源的开发和利用，随着社会的发展，越来越多的数据资料不断积累，从这些数据中挖掘规律和知识，对于更好地开发和利用各种自然资源具有重要的意义。欧洲经济共同体一位负责人曾经说过一句非常深刻的话："欧洲的许多矿产资源存在于数据库中"，我们可以这样理解：通过矿产数据的分析，揭示在数据中存在的信息，可以对矿产赋存特点有进一步的认识，进而帮助我们找到更多的资源。

无论是作为勘探地质工作者，还是作为油气藏工程师，都越来越强调对地质现象或地质体定量研究的能力，尤其是随着资料的不断积累，这种趋势在不断加强。

利用数据资料形成认识的过程是一个分析和归纳的过程，而利用这种认识指导实践的过程是一个推断过程，前者是途径，后者是目的，认识世界最终是为了改造世界。

定量分析的基础是数据，分析过程是建模过程，建模有确定型和随机型两种模式。确定型模式探求的是某一变量与其影响因素之间的函数关系，这当然是科学研究应当追求的最终模式，因为，函数关系具有逻辑和形式上的美学意义，它带来的是精确的预测，是人们对研究变量百分之百的掌控。然而，自然现象不同于人类可以操控的室内物理实验，任何一个自然变量都是在"上帝"操控下的实验结果，尤其是地质变量，不仅有"无数只手"在影响它、改变它，而且，这种影响和改变历经漫长岁月，使得变量与其影响因素之间的关系变得极其复杂，以至于建立函数关系的企图遭遇严峻的挑战。

然而，复杂中孕育着简单，大量的数据表现出规律，也正是如此，统计学这一基本的工具在地质现象的研究中得到了广泛的应用。显然，用统计学这一工具研究地质现象，体现的正是人类的智慧和不屈不挠的科学精神，数学地质也正是在这一基础上不断发展成为一个庞大的知识系统。

现代数学地质涉及地质学的诸多领域，对于油气地质工作者来说，究竟需要掌握哪些内容，也就是说，一本适合的石油高等院校教材应包括哪些主要内容？这是我们长期思考的一个问题。在中国石油大学的教学工作为我们提供了一条探索这一问题的有效途径。

我们从 1994 年开始，先后在中国石油大学（北京）和中国石油大学（华东）主讲了研究生（包括进修生）和博士生课程《地质统计学》、《分形几何》、《数学地质》和《现代数学地质》，累计 20 次之多，对研究生层次的数学地质培养进行了有益的探索，在空间相关统计学和分形几何方面，进行了理论方法和应用实例方面的总结，同时，相关方面的科研工作也起到了一定的推动作用，使这部分内容逐步得到了充实和完善。

本科生数学地质的教学实践使我们回过头来，考虑更加基础的内容——概率论、数理统计、线性回归和多元统计分析。教材在体现这些内容时试图从简单的实例出发引出定理和公式。

希望这本教材能引导更多的读者对数学地质产生兴趣并在今后的工作中自觉地应用数学地质方法，促进研究工作向半定量化和定量化方向发展，为油气勘探开发事业做出更大的贡献。

教材的编写和顺利出版是多方面持续支持和努力的结果，我们首先要感谢中国石油大学

(北京)资源与信息学院书记庞雄奇教授、院长朱筱敏教授和副院长陈小宏教授和钟宁宁教授对该教材出版的大力支持。我们感谢资源与环境系和盆地与油藏研究中心的领导与同事们对我们教学工作的支持和帮助。同时,我们要特别感谢彭仕宓教授对教材编写工作的支持和尹志军博士对教材编写的贡献。

中国石油大学张一伟教授和熊琦华教授长期致力于统计学方法在油藏定量描述中的应用,有力地推动了现代数学地质在石油大学的教学和科研应用工作,本教材也凝聚着前辈的心血,在此,对他们的杰出工作表示崇高的敬意。这里还要提及的是金之钧教授首先提出在中国石油大学(北京)为研究生开设数学地质课程的计划并编写了教学大纲,虽然因其有其它更重要的工作而未讲授该课程,但他的努力为我们后来开设数学地质课程奠定了基础,也为本教材的形成提供了教学实践契机,在此表示感谢。

本教材前言、绪论、第一章、第二章的第一节到第四节中的参数估计部分以及第五章到第七章由康永尚编写,其中,第七章中随机模拟应用实例由尹志军博士提供;第八章以康永尚编写为主,沈金松补充了部分应用实例;第二章第四节中的假设检验部分以及第三章到第四章由沈金松编写,在第四章编写过程中吸收了旅居加拿大的谌卓恒博士提供的素材和应用实例;全书由康永尚统稿。研究生杨帆、徐显生和徐桂华等协助完成了教材中部分图件的清绘和数据表的制作及其它事务性工作,对他们的工作表示感谢。

中国石油大学(北京)陈安乐教授对全书进行了详细的审校,提出了大量宝贵意见,在此,编著者衷心感谢陈老师的辛勤劳动。

本教材是编著者长期教学实践的总结和凝练,编著者的家庭给予的支持是完成这一工作最有力的保证,在此表示感谢。

尽管几经修改,但教材中的错误在所难免,欢迎并感谢读者能给予指正并与我们取得联系,以便该教材体系能在未来的实践中不断完善。我们的联系方式是:kangysh@sina.com,jinsongshen@hotmail.com,zchen@shaw.ca。

编著者

2005 年 4 月

目　　录

绪　论

《现代数学地质》学习目的和要求：能够把概率论和数理统计知识熟练地应用到油气勘探开发的决策分析和地质变量变化规律的研究和预测中，掌握多元统计学的基本原理及其应用，深刻理解空间相关统计学的精髓并能应用于空间变量的定量化研究中，了解非线性科学的新思维和分形几何的基本原理及其应用。

一、数学地质的起源与发展

1.数学地质的定义

关于数学地质的定义，尚未有一致的说法，Davis(1973)把数学地质定义为"地质数据的定量分析方法"，而 Agterberg(1974)把数学地质定义为"地球科学中的全部数学应用"，赵鹏大(1983)认为"数学地质是研究最优数学模型并查明地质运动数量规律性的科学"。无论采用何种定义，我们都可以把数学地质理解为以数学作为基本的理论和方法，以计算机作为工具，对地质现象进行定量化研究的一门边缘学科，是由地质学、数学和计算机科学互相结合而逐渐发展完善的。

2.数学地质产生的原因及对地球科学的影响

传统的地质学基本上是一门描述性的学科，从事地质学研究的人对数学在地质学研究中的应用及所起的作用还存在不同的看法。然而，随着人类对地球研究和探索的深入，需要研究多种地质因素对某一地质现象的影响，分析多种地质因素间的相互关系及各自的相对重要性，对于这样的一系列复杂问题没有数学作为工具是不行的。其次，随着研究手段的多样化，通过实验和观察获得了大量的数据，要利用这些数据研究各种量之间的关系及各种量的变化等反映地质运动和变化的规律，必须应用数学方法和计算工具。再次，数学是研究和模拟各种地质现象的理论基础，大气运动规律、地壳变化和运动的研究等都需要数学模拟。

随着数学向传统地质学的渗透和应用，地质学研究逐渐由定性向定量分析过渡，也由最初的单变量分析向多变量分析过渡。所研究的对象也由确定性事件向随机性事件过渡，对地质事件和地质现象的分析也由定性解释向定量分析与数值模拟相结合的阶段发展。

3.数学地质发展的 4 个阶段

数学被逐步引入到地质学中经历了 160 多年的历史，作为一门新兴的边缘学科，数学地质已有 60 多年的发展历史。从应用的广泛性看，数学地质的发展大致经历了以下 4 个阶段。

1)孕育阶段(1950 年以前)

1840 年英国地质学家 Lyell 首次以古生物化石的统计分析为论据，对第三纪地层进行了划分，确定了岩石地层次序，著述了《定量动物学》一书，开创了数学方法引入地质学问题研究的先例。上世纪 30 年代后期，Simpson(1939) 著述的《分析地质学》一书中列举了统计学在生物研究中的多方面应用。Burma(1949)在"多元分析——地质学和古生物学中的一种模型分析工具"一文中明确提出了多元统计方法是一种最有前景的生物计量方法。Krumbein 和格里菲斯在沉积学研究中也使用了概率统计方法，为用数学方法研究地质现象和个别指标的统计分析解决具体地质问题奠定了基础。这一阶段也是统计分析在地质学中应用的可能性问题讨论最

为激烈的时期。

2)早期阶段(1951~1960)

Krumbein(1956)在研究岩石的矿物成分、岩性和化学成分时,应用了多元统计方法,并把岩石成分作为 n 维空间中的一个点或向量进行统计处理。尤其值得指出的是,1958 年由 Krumbein 和斯洛斯公开发表了第一个面向地质应用的计算机程序,标志着计算机技术在地质研究中应用的开端,加速了地质学的定量化研究进程。1958 年 Sichel 和 Kridge 编著的《地质统计学》及 Allais(1957)发表的《单元中矿床数服从泊松分布的矿产资源定量评价》等重要文献奠定了地质统计学分支学科的基础,是数学地质早期发展的重要阶段。

3)形成发展阶段(1961~1980)

自 1968 年在布拉格第 23 届国际地质大会上成立国际数学地质协会(IAMG)并由维斯捷列马斯任第一届国际数学地质协会主席以来,数学地质这一边缘交叉学科得到了长足的发展。1969 年出版了两种专门的数学地质刊物——《Mathematical Geology》和《Computers & Geosciences》,后来又设立了新刊物《Nonnewreable Resources》。随着人类对资源需求的增加,社会对资源评价和预测研究精度的要求也逐渐提高,迫使地质学家和工程技术人员寻求更加可靠和精确的评价和预测模型。例如,在石油禁运和社会对油气需求增大的背景下,Harris(1973)发表了多元统计评价及主观概率评价两种油气资源评价模型。为了加强矿产资源的定量评价,Agterberg(1974)针对法国和非洲的一些固体矿床的储量和品位的评价,提出了矿产资源评价的逻辑模型。

4)深入广泛的发展阶段(1980 年至今)

这一阶段中,数学地质向更广泛和更高水平发展。随着超大规模集成电路的计算机的研制成功和各种新的应用数学模型的建立,促进了数学地质向更加深入和稳健的方向发展。如日本林知已夫的数量化理论,美国麦克卡门和波特波尔的特征分析和前苏联的康士坦丁诺夫的逻辑信息方法等为定性地质变量和定量地质变量的联合数学模型的建立提供了基础。地质过程的计算机模拟、地质数据库的建立和地学领域内人工智能专家系统的研制和应用,为数学地质的广泛应用提供了更加先进和方便的手段。

由于计算机和计算数学的发展,新的多元统计分析方法和新的数学理论和方法不断与地质结合,使数学地质更加完善和成熟。人们开始研究一些新的数学地质方法,如法国著名的数学家 Tom(1968)提出了用数学工具描述灾难性或突如其来的变化现象的突变理论,后来又将微分拓扑学的研究成果应用于地质学,建立了地质学中的突变理论模型,用于研究断层运动、二叠纪海洋无脊椎动物的灭绝原因等。

人工智能及专家系统的研究使专家的知识和经验能为他人所用,充分发挥了计算机和专家的作用;自美国斯坦福国际人工智能研究所于 1976 年研制成功地学领域内的第一个专家系统——Prospector 以来,利用该系统美国的一家勘探公司已在华盛顿州发现了一个钼矿床。美国的石油资源评价专家系统及斯仑贝谢跨国测井公司研制的地层倾角测井资料处理解释咨询系统(Dipmeter Advisor)已在油气勘探和开发中取得了很好的效果。地质数据库的成功应用为资料的交流和综合应用提供了可能。

二、数学地质的最新进展

从近年数学地质发展的趋势看,数学地质正全方位地向各地质分支学科渗透,数学地质与新发展的其它学科交叉的深度和广度在不断扩大,较为突出的表现为以下几个方面:

(1)非线性科学在地质学中的应用。

自变量的连续改变而发生不连续变化的数学理论——突变理论的发展。突变现象的共同特点是，条件微变会导致系统宏观状态的剧变。例如，地质学中的突变理论模型；用于研究断层运动、二叠纪海洋无脊椎动物的灭绝原因等方面取得了可喜的进展。

(2)人工智能在地质学中的应用：专家系统在地质学中的应用；多种人工智能技术在地质学中的综合应用，如神经网络、遗传算法等。

(3)地理信息系统：在计算机软件的支持下，空间数据输入、存储、检索、运算、显示和综合分析的应用技术系统在地学领域中的应用。

(4)数据库技术在地质学中的应用。

(5)自动绘图技术在地质学中的应用。

(6)计算机的最新硬软件在地质学中的应用。

从数学地质的最新进展来看，"非线性科学"认识论在理论分析方面的开拓和已成熟方法的软件化，代表两个主流。

三、现代数学地质包含的主要内容

根据以上对数学地质定义的理解，我们可以说数学地质是一门迅速发展、开放的学科。随着计算机技术的不断提高和地球科学的发展，计算机应用的日益普及，地质现象的定量化研究已成为一个必然趋势，数学地质要解决的问题正是地质现象的定量化这一极为困难的问题。作为一门把地质问题定量化不断推向深度和广度的、具有生命力的综合性边缘学科，正蓬勃发展着，它以多种数学方法为基础，同时吸收了现代物理学、化学、社会科学等研究成果，借助于计算机，实现对地质过程的模拟、对地质现象时空变化的分析、对多种变量之间相互关系的刻画，而以各种预测为最终目标，为人类的生产实践提供决策的量化依据。《现代数学地质》包括以下主要内容：

(1)数据分布类型及预处理：地质数据的分布类型及其检验数据的特征分析，变量的选择及最佳组合，地质数据之预处理；

(2)属性分析与综合：相关、聚类、判别、对应、因子分析等；

(3)地质多元统计分析方法：趋势面分析、特征的空间分布、单位向量场分析；

(4)矿产资源统计预测：Delphi 法、主观概率、蒙特卡罗、条件概率等；

(5)时间序列分析预测；

(6)地质统计学；

(7)模糊数学和突变论；

(8)分形理论；

(9)灰色系统；

(10)地质过程的计算机模拟(如盆地模拟)；

(11)地质数据库及地质数据处理系统(如 GIS)；

(12)人工智能专家系统。

现代数学地质是一个庞大的知识体系，根据我们的教学经验，针对油气勘探开发地质和工程人才培养的需要，试图建立起从数理统计基础到专业应用的一个实用且相对完整的知识结构体系并努力使本教材具有如下特色：

(1)针对性：主要针对石油勘探开发的需求设计教材的内容。

(2)新颖性：涉及风险分析、空间相关统计学、非线性科学思维观和分形几何等方法，有些方法在国内目前应用还比较少，但随着我国社会经济的发展、资料数据的积累尤其是油公司体制的逐步确立，其应用前景良好。

(3)系统性：从概率论基础开始，逐步深入，过渡到空间相关统计学比较前沿的领域，最后，引入非线性科学的思维观和方法论。

(4)学科交叉性：软硬学科交叉(决策科学和地质变量定量表征)、勘探和开发的学科交叉。

(5)实用性：每部分方法原理之后，附有在勘探开发领域内的应用实例。

(6)应用的广泛性：适用于"地质工程"、"矿产普查与勘探"、"地球探测与信息技术"、"油气井工程"、"油气田开发工程"和"环境科学"等多个二级学科。

根据以上思路本教材设计了8章内容，从概率论和数理统计的基础知识到其应用，从单变量到多变量，从空间无关的变量到空间相关的变量，从线性科学的思维观到非线性科学的思维观，从易到难渐次深入：第一、第二章作为概率论和数理统计的基础知识，从基本概念、基本公式、基本方法和在油气勘探开发中的应用几个方面进行了介绍和讨论；第三章为线性相关和线性回归，其重点是阐明最小二乘法的基本原理及简单和多元线性回归方程的建立与应用；第四章为多元统计学，主要介绍了趋势面分析、主成分分析、聚类分析和判别分析等多变量研究中常用的方法；第五章到第七章为空间相关地质变量研究——地质统计学的主要内容，从区域化变量的结构分析、普通克立金、指示克立金和地质变量的随机模拟及其应用等方面进行了分析讨论；第八章涉及非线性科学的基本思维观和分形几何的基本原理及其应用。

第一章　概率论基础及其在油气勘探开发决策中的应用

本章学习目的和要求：简要回顾概率论基本知识，介绍概率论在油气勘探风险决策分析中的应用，为以后各章的学习奠定基础，并使学生能熟练地把概率论的基本方法应用到油气勘探开发的决策分析中。

第一节　概率论基础

一、基本概念

在自然界和人类社会中存在着两类不同的现象，一类是确定性现象，另一类是随机性现象。

1.确定性现象和随机性现象

在一定条件 S 下进行某种试验或观察，事先可以预料必然出现某种结果，或者说在条件 S 不变的情况下，重复进行一系列这种试验或观察，所得结果是完全相同的。像这类在一定条件下必然出现某种结果的现象称为确法性现象，如纯水在一个大气压下加热到 100℃时必然沸腾，异性电荷必然互相吸引、同性电荷必然互相排斥等皆为确定性现象。

在一定条件 S 下进行某种试验或观察，在试验或观察之前无法预知确切的结果，只知道可能出现这样的结果，也可能出现那样的结果。当然，一旦试验或观察进行之后，出现什么结果就是完全确定的了。但是，在试验条件 S 不变的情况下，重复进行一系列这种试验或观察所得的结果却是不尽相同的。像这类在一定条件下可能出现的结果不止一个，至于出现哪一个，事先又无法确定的现象称为随机性现象。如某一地区在某一天的降雨量，电话交换台在单位时间内收到的呼唤次数等皆为随机性现象。

2.随机试验和样本空间

自然界和人类社会中各种各样的现象常常在人们所进行的试验或观察中呈现出来，我们把呈现出随机性现象的试验或观察称为随机试验，记为 E。

在进行一个随机试验时，试验的所有可能结果应该是明确知道的，每一个结果被称为样本点，用 ω 表示。由样本点 ω 的全体构成的集合称为样本空间，用 Ω 来表示，记为 $\Omega=\{\omega\}$。

【例 1-1】　掷一枚骰子，试验的所有可能结果有 6 个，即 $\omega_1=$“1”，$\omega_2=$“2”，$\omega_3=$“3”，$\omega_4=$“4”，$\omega_5=$“5”，$\omega_6=$“6”，则样本空间为

$$\Omega=\{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5,\omega_6\}$$

【例 1-2】　观察电话交换台在单位时间内收到的呼唤次数，若用 k 表示“在单位时间内收到 k 次呼唤”这一结果，$k=0,1,2,\cdots$，则样本空间为

$$\Omega=\{0,1,2,\cdots\}$$

3.随机事件

在一个随机试验中,对一次试验而言,可能出现也可能不出现的事情,称为随机事件,记为 A。对一个随机试验来说,显然,它的每一个可能的结果(即每个样本点)都是一个随机事件,我们称之为基本事件,像例 1-1 中出现 6 点的事件即为一个基本事件。在一个随机试验中,除了基本事件外,还有由若干个可能结果(即若干个样本点)所组成的事件,相对于基本事件,称这种事件为复合事件,如例 1-1 中出现偶数点的事件即为一个复合事件。

4.频率与概率

在相同条件 S 下,重复进行 n 次试验,若在 n 次试验中,事件 A 发生的次数(可称为频数)为 μ_A,则比值 $\frac{\mu_A}{n}$ 称为事件 A 在 n 次试验中发生的频率 f,即

$$f = \frac{\mu_A}{n} \tag{1-1}$$

如果当 n 充分大时,A 发生的频率 $\frac{\mu_A}{n}$ 稳定地在某一数值 p 附近摆动,而且一般来说随着试验次数的增多,这种摆动的幅度越变越小,即频率 $\frac{\mu_A}{n}$ 越来越稳定于 p,则称 p 为此随机试验中随机事件 A 发生的概率,记作

$$P(A) = p \tag{1-2}$$

【例 1-3】 表 1-1 为 17 块岩心的实测岩石密度数据,假设事件:$A_1 = [2.65, 2.68]$;$A_2 = [2.68, 2.71]$;$A_3 = [2.65, 2.71]$;$A_4 = [2.66, 2.69]$,计算各事件的概率。

解: $P(A_1) = \frac{7}{17} = 0.412$;$P(A_2) = \frac{9}{17} = 0.529$;$P(A_3) = \frac{16}{17} = 0.941$;

$P(A_4) = \frac{12}{17} = 0.706$

表 1-1 实测岩石密度数据表

岩心号	岩石密度 ρ, g/cm³	岩心号	岩石密度 ρ, g/cm³	岩心号	岩石密度 ρ, g/cm³
1	2.68	7	2.69	13	2.69
2	2.68	8	2.70	14	2.68
3	2.68	9	2.68	15	2.68
4	2.69	10	2.69	16	2.68
5	2.69	11	2.70	17	2.70
6	2.70	12	2.74		

5.事件之间的关系与运算

1)包含关系

若事件 A 发生时,事件 B 必发生,则称事件 B 包含事件 A,见图 1-1(a),记为 $A \subseteq B$。例 1-3中的 A_3 包含事件 A_1,同时,A_3 也包含事件 A_2。

2)事件的和

若某事件发生当且仅当事件 A 与事件 B 中至少有一个发生(或者说当且仅当事件 A 发生或事件 B 发生),则称此事件为事件 A 与事件 B 的和见图 1-1(b),记为 $A + B$ 或 $A \cup B$。

3)事件的积

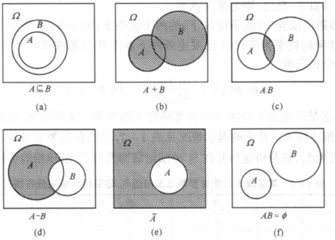


图 1-1 事件之间关系及运算示意图

若某事件发生当且仅当事件 A 与事件 B 同时发生，则称此事件为事件 A 与事件 B 的积，见图 1-1(c)，记为 AB 或 $A \cap B$。例 1-3 中事件 A_1 和 A_4 的积(同时发生)为事件[2.66,2.68]。

4)事件的差

若某事件发生当且仅当事件 A 发生而事件 B 不发生，则称此事件为事件 A 与事件 B 的差，见图 1-1(d)，记为 $A-B$。

5)对立事件

事件“A 不发生” 为事件 A 的对立事件，记为 $\bar{A}$，见图 1-1(e)。

6)互不相容事件

若事件 A 与事件 B 不能同时发生，则称事件 A 与事件 B 为互不相容事件，见图 1-1(f)。例 1-3 中 A_1 和 A_2 两个事件即为互不相容事件。

6. 概率的可加性

1)概率的有限可加性

若事件 $A_1, A_2, \cdots, A_n$ 两两互不相容，则

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) \tag{1-3a}$$

式(1-3a)也可写成如下形式

$$P(\sum_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \tag{1-3b}$$

2)概率的可列可加性

若事件 $A_1, A_2, \cdots, A_n \cdots$ 两两互不相容，则

$$P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \tag{1-4}$$

二、条件概率公式和概率的乘法公式

1. 条件概率公式

随机事件 A 的概率是在条件 S 之下来考虑的，若除条件 S 之外，不再附加任何条件，则称

这种概率为无条件概率，反之，则为条件概率。

设 A、B 是某随机试验中(或条件 S 下)的两个事件 $P(A)\neq 0$。称事件 B 在“事件 A 已发生”这一附加条件下的概率为事件 A 已发生条件下事件 B 的条件概率，记 $P(B/A)$。条件概率可由条件概率公式求得

$$P(B/A) = \frac{P(AB)}{P(A)} \qquad P(A) > 0 \tag{1-5}$$

【例 1-4】 在某区块钻 5 口井，5 口井都钻穿了侏罗系三工河组 S_2 砂层组，该砂层顶部、中部和底部的沉积相在 5 口井中的揭示情况见表 1-2，现正在钻第 6 口井，已钻到 S_2 砂层组的中部，其沉积相为 F_1，试分析这口井在 S_2 砂层组底部发育 F_3 沉积相的可能性有多大？

表 1-2 某区块 5 口井侏罗系三工河组 S_2 砂层组沉积相分布情况

位置	井$_1$	井$_2$	井$_3$	井$_4$	井$_5$
顶部	F_3	F_2	F_1	F_2	F_1
中部	F_1	F_1	F_2	F_1	F_2
底部	F_2	F_3	F_3	F_3	F_3

解:设事件 A = [F_1 是 S_2 砂层组中部的沉积相]，事件 B = [F_3 是 S_2 砂层组底部的沉积相]，那么，要回答的问题是当事件 A 发生时，事件 B 发生的条件概率。

根据表 1-2 的信息，我们有

$$P(A) = \frac{3}{5}, P(B) = \frac{4}{5}$$

事件共同发生的概率为 $P(AB) = \frac{2}{5}$，则在事件 A 已发生的条件下事件 B 发生的概率为

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{2/5}{3/5} = \frac{2}{3}$$

故有：当第 6 口井 S_2 砂层组的中部沉积相为 F_1 时，在 S_2 砂层组底部发育 F_3 沉积相的可能性为 2/3。

2. 概率的乘法公式

由式(1-5)可得概率的乘法公式

$$P(AB) = P(A)P(B/A) \tag{1-6}$$

在关系式 $P(AB) = P(A)P(B)$ 成立的前提下，称事件 A 和 B 相互独立。

注意，在讨论两个事件 A 和 B 的独立性时，A 和 B 必须是相容事件，否则，就无从谈其独立与否。

在例 1-4 中，$P(AB) = \frac{2}{5} \neq P(A)P(B) = \frac{12}{25}$，故 A 和 B 两个事件不是相互独立的，也就是说，在 A 发生时，对 B 发生的概率会产生影响。

三、全概率公式和概率树

1. 全概率公式

在实际问题中，我们常常遇到这样的事件 B，它总是与若干个两两互不相容的事件 A_1，$A_2 \cdots A_n$ 之一同时发生。这时，为了计算事件 B 的概率，可先把事件 B 分解为一些不相容的简单事件之和，即 $B = BA_1 + BA_2 + \cdots + BA_n$，然后分别计算和式中的简单事件的概率，再利用概率的可加性，最终得到 $P(B)$。下面，先从实例分析入手，引出全概率公式。

【例 1-5】 某油层中有三种沉积微相,已知其中第一、第二和第三种沉积微相各占总量的 50%、25%和 25%,又知第一、第二和第三种沉积微相高渗透($>100\times10^{-3}\mu m^2$)岩心出现的概率分别为 1%、2%和 4%,现任取一个岩心,问取出的岩心具有高渗透率的概率有多大?

解:设 A_i="取出的岩心是第 i 种沉积微相",则有

$$P(A_1)=50\%,\quad P(A_2)=25\%,P(A_3)=25\%$$

再设 B="取出的岩心是高渗透率",事件 B 总是与三个不相容的事件 A_1、A_2 和 A_3 之一同时发生,且 $P(B/A_1)=1\%$,$P(B/A_2)=2\%$,$P(B/A_3)=4\%$,根据全概率公式,事件 B 的概率为

$$P(B)=\sum_{i=1}^{3}P(A_iB)=\sum_{i=1}^{3}P(A_i)P(B/A_i)$$
$$=50\%\times1\%\times25\%\times2\%\times25\%\times4\%=2\%$$

显然,事件 C="取出的岩心不是高渗透率"的概率为

$$P(C)=1-P(B)=1-2\%=98\%$$

由以上实例分析,可引出全概率公式如下:

设:(1)$A_1,A_2,\cdots,A_n$ 两两互不相容,$P(A_i)>0\quad(i=1,2,\cdots,n)$

(2)$B\subseteq\sum_{i=1}^{n}A_i$(事件 B 包含在 $\sum_{i=1}^{n}A_i$ 中)

则有

$$P(B)=\sum_{i=1}^{n}P(A_i)P(B/A_i)=\sum_{i=1}^{n}P(A_iB)\tag{1-7}$$

2.概率树

我们可以用树形结构图清晰地表示事件之间的概率关系,例 1-5 中事件之间的概率关系可用图 1-2 所示的概率树来表示。

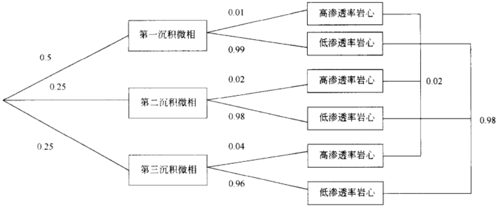


图 1-2 例 1-5 中事件之间概率关系的概率树

四、贝叶斯公式和逆概率树

1.贝叶斯公式

在实际工作中还常常碰到这样一类问题:已知某个试验结果是由许多原因导致的,如果人们通过试验观察到这个结果,于是人们希望通过这个信息来探讨(反求)每个"原因"导致这个结果的可能性有多大,如例 1-4 中,若已知取出的一个样品为高渗透样品,问这个样品属于第

一、第二和第三种沉积微相的可能性有多大？由此引出贝叶斯问题，贝叶斯问题实际上是逆概率问题，在决策过程中有十分重要的应用价值，贝叶斯公式由下列方式表述：

若有一组事件 $A_1, A_2 \cdots A_n$ 以及事件 B，它们满足公式(1-7)中相同的条件，而且 $P(B) > 0$，则有贝叶斯公式

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{\sum_{j=1}^{n} P(A_j)P(B/A_j)} = \frac{P(A_iB)}{P(B)} \quad (j = 1,2,\cdots,n; \quad i = 1,2,\cdots,n) \quad (1-8)$$

实际上，式(1-8)的贝叶斯公式可根据条件概率公式和全概率公式直接推导出来。

由式(1-8)所计算的概率称为后验概率，它实质上是 B 发生条件下 A_i 发生的条件概率，这不同于 A_i 本身发生的概率，称为先验概率。

下面，用一个例子来说明贝叶斯公式的应用。

【例 1-6】 回到刚刚提出的问题上，即在例 1-5 中，若已知取出的一个岩心样品为高渗透率样品，问这个样品属于第一、第二和第三种沉积微相的可能性各有多大？

解：由贝叶斯公式可计算取出的一个样品为高渗透率样品(事件 B)时，该样品为第一、第二和第三种沉积微相的概率分别为

$$P(A_1/B) = \frac{50\% \times 1\%}{2\%} = 25\%$$

$$P(A_2/B) = \frac{25\% \times 2\%}{2\%} = 25\%$$

$$P(A_3/B) = \frac{25\% \times 4\%}{2\%} = 50\%$$

故有：当已知取出的一个样品为高渗透率样品时，这个样品属于第一、第二和第三种沉积微相的可能性分别为 25%、25% 和 50%。

2. *逆概率树*

贝叶斯公式可用逆概率树清晰地图示表达，例 1-6 中的逆概率问题可用图 1-3 中的逆概率树来表达，为了构建图 1-2 中的逆概率树，还需要计算出已知取出的一个样品为低渗透样品(事件 C)时，这个样品属于各沉积微相的可能性，即

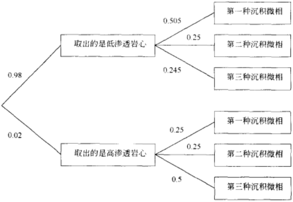


图 1-3 逆概率树

$$P(A_1/C) = \frac{50\% \times 99\%}{0.98} = 50.5\%$$

$$P(A_2/C) = \frac{25\% \times 98\%}{0.98} = 25\%$$

$$P(A_3/C) = \frac{25\% \times 96\%}{0.98} = 24.5\%$$

第二节　概率论在油气勘探风险分析和圈闭期望价值计算中的应用

一、油气勘探中的风险和成功概率

1.油气勘探中的风险

无论在哪个经济领域进行投资，风险因素都是投资者所必须考虑的因素之一。油气勘探是寻找、发现圈闭，评价其经济可行性(风险性)。钻探目标，直至发现商业性油气藏的过程，是一个高投资、高风险、高收益的领域。油气勘探的风险包括地质风险、工程风险、经济风险和政治风险等，其中，地质风险是最根本的。

地质风险涉及有无油气及有多少油气的问题，石油深埋在地下几百米到几千米，其生成、运移和保存等过程，都有许多不确定性，虽然人们的认识在不断加深，且科学技术在不断进步，但这种认识与实践、预测与实际之间仍常常发生偏离，这就存在着风险的可能性，如表 1-3 为 AAPG 钻井资料统计委员会(CSD)收集的美国钻井统计资料，说明了各类井的历史统计成功率，新区野猫井的勘探成功率只有 13% ~ 18%。

表 1-3　CSD 报道的 20 世纪 80 年代美国陆上和海上不同类型井的成功率

井类型	成功率	井类型	成功率	井类型	成功率
开发井	75% ~ 80%	探边井(评价井)	40% ~ 45%	新区野猫井	13% ~ 18%
所有探井	20% ~ 30%	老区野猫井	25% ~ 35%		

2.成功概率的定义

对一个给定的圈闭，地质学家能够估计出关键地质因素(储集岩、烃类充注等)存在的概率值，这些地质因素被认为是地下油气聚集的必要条件，这就引出了烃类存在的概率或地质成功率的估计问题。

我们常常会碰到这样一些问题：回收完井成本及后续生产操作费用的成功率有多大？同时回收钻井费用的成功率有多大？获得合理利润的概率有多大？因此，成功概率有许多种定义，用来有效地预测成功概率的技术是存在的，但应用这些技术时需要使用关于成功概率的一致的、严格的定义(图 1-4)。

(1)地质成功井——钻遇可流动的、可计量的烃类的井。

(2)完井成功井——可完井投入生产的井。

(3)边际成功井——忽略探井钻探成本(作为沉没成本)，仅考虑完井和生产操作成本才能够盈利的井。

(4)商业成功井——发现的油气能够负担开发井、完井、地面设施、生产操作成本、井口税，

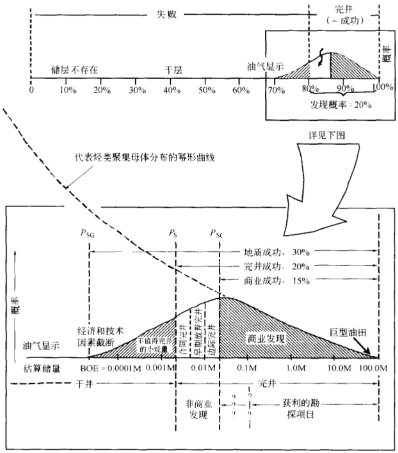


图 1-4　不同成功概率定义示意图

图中 P_{SG}表示地质成功率，P_S表示完井成功率，P_{SC}表示商业成功率

并能在回收相应数目的干井成本的基础上获得合理的利润的发现井。注意，这个定义容许包括一小部分井，这部分井的发现不能负担其它的勘探成本，如地质和地球物理(G&G)勘探成本、矿区租金以及相应数目的干井成本。

3.地质机会因子和成功概率

对于所有的勘探圈闭，地质勘探人员都应该独立地给出任何一个圈闭的4个重要地质机会因子存在的概率。

1)有效储集层

孔隙度和渗透率达到生产标准，最小厚度足以容纳可计量的、可流动的油气，或足以促使审慎的国内陆上勘探者冒险完井的储集层存在的概率有多大?

一个公司用油气的可流动性，以满足某种最低标准的需要。在任何情况下，我们所寻求的是地质人员关于一个最小规模储集层存在的置信度——厚度、孔隙度和有效渗透率。在此情况下，钻遇水层但却是一个具有商业质量的砂岩储层就储集层来说不算是失败，但从另外几方面来说，如圈闭幅度低、缺乏烃类充注或圈闭漏失，钻探则是失败的。然而，在预测有10英尺孔隙砂岩存在的目的层钻遇的是1英尺的致密砂岩，就储集层的角度来说钻探则是失败的。

2)构造落实程度

平面图和剖面图上展示的目的层的构造能够代表地下真实构造的概率有多大？在这里并不是指圈闭本身是否存在，而是指圈闭平面图和剖面图描述地下真实情况的准确程度。

这个地质机会因子常常应用于地层和构造圈闭预测，而且毫无疑问在大多数勘探和开发工程中，构造图通常都是惟一最重要的图件。同样，构造解释的偏差也是造成干井的普遍原因。

地质构造概率因子结合储层条件，主要用于描述设想存在的油藏或气藏的几何形态和流体体积，而这两方面对于进行生产试验或审慎的钻杆测试都是必要的。

3)烃类物质基础

地下地质环境中存在烃而且圈闭已经接收到一定量烃类的概率有多大？这个地质机会因子描述诸如源岩体积是否足够大、是否有油和/或气生成、油气到圈闭的运移通道以及储集层流体中烃类饱和度的问题，时间匹配问题此处不予论述。在大多数新近展开勘探的盆地中，对烃类充注的置信度相对较低，而在勘探程度较高的盆地中和正在开发的区带中，对烃类充注的置信度是比较高的。

4)时间匹配

根据所描述的岩性组合和构造形态，圈闭存在以及油气运移到圈闭区域时圈闭已经形成的概率有多大？这里我们讨论三个问题，首先是储集层与顶、底和侧向封隔层之间的封闭性（不论是地层成因还是断层成因的），流体粘度、地层厚度、差异渗透率和断层活动历史都影响着封闭性。其次是前面第三条中提到的时间匹配问题：如果圈闭在油气运移之后才形成，那么与"等马都跑出后才把门关上"没什么区别了。第三个问题是后期淡水冲刷和储集体内的烃发生降解带来的保存问题。这里所用的术语"圈闭"没有任何几何形状和构造形态的含义——只有容量和封闭性，圈闭充满程度这个棘手的问题就属于这个讨论范畴，最好用百分数来描述。对于大多数开发井，圈闭封闭条件都已经满足。

把以上4个地质机会因子都看作具有概率范围为0～1的独立变量，4个因子连乘得到油气藏存在的概率。在美国陆上探区中，根据以上定义的这4个地质因素的4个概率值的连乘积与完井成功概率非常接近，把它记为成功概率 P_s，由此可以得到失败的概率 P_f，此处 $1.0 - P_s = P_f$。因此从定义来说，P_s 与完井成功概率非常一致，如图1－4所示。

二、圈闭期望价值的计算

1.圈闭期望价值的概念

油气勘探风险项目的经济分析必须在这样的假设前提下进行：即项目是成功的，这种成功通常被表述为盈利性的几个等级之一，而盈利性是由影响项目商业性的地质技术经济参数的不同范围确定的。例如，一个机构可以通过预测的储量规模的最小值、众数、中值、平均值和最大值进行经济评价（注意圈闭储量分布最好的单个代表值是平均值）。许多勘探项目失败了，这种失败的后果在评价钻探圈闭的经济价值时应该考虑进去。期望价值（*EV*）既包括了钻井成功与失败的概率，又包括了成功与失败所对应的两种经济后果：

$$EV = P_s \times (\text{项目净现值}) + P_f \times (\text{失败成本}) \tag{1-9}$$

式中，P_s 和 P_f 分别对应成功和失败的概率。

面对多种选择，而且每种选择的期望价值均为正值，决策原则为选择期望价值最高的选项。很明显，如果有人选择了期望价值为负值的风险项目，那他就是在赌博。

在勘探风险项目中，圈闭的净现值（NPV）的计算通常要考虑所有的前期勘探成本。勘探失败的成本通常包括干井成本、租金和地质、地球物理勘探成本。对于特别大的圈闭还要包括一口或多口探边井的成本。

2．圈闭期望价值的应用

期望价值在油气勘探风险项目分析中是很有用的概念，它使我们能够对不同圈闭的价值进行权衡。通过它，可以在具有较大储量潜力但成功概率较小和储量潜力适中但成功概率较大的圈闭之间进行对比。例如，在表1－4中只需通过期望价值（EV）就可以判断圈闭 A 比圈闭 B 好，其原因是圈闭 A 具有较高的期望价值。然而，决策者实际上可能会选择圈闭 B，以较高的成功概率和较低的前期成本换取较少的回报。即使这样，期望价值在确定和处理这个例子中的风险规避行为时还是很有用处的。

表1－4　两个圈闭期望价值的计算与对比

圈　闭	计　算　方　法
圈闭 A	10000 桶平均储量扣除所得税后的估计现值＝30,000,000 \$ 估计的成功概率＝10%，失败概率＝90% 扣除所得税后的干井成本估计值＝－2,500,000 \$ EV＝0.10（＋30,000,000 \$）＋0.90（－2,500,000 \$）＝＋750,000 \$
圈闭 B	10000 桶平均储量扣除所得税后的估计现值＝30,000,000 \$ 估计的成功概率＝40%，失败概率＝60% 扣除所得税后干井成本的估计值＝－1,000,000 \$ EV＝0.40（＋30,000,000 \$）＋0.60（－1,000,000 \$）＝＋600,000 \$

圈闭排队表在预测勘探部署总的结果中具有相当大的价值，它的使用可融合到各级勘探部门及整个油公司的勘探部署中，该表的建立有赖于圈闭期望价值的确定。决策树分析也有赖于可靠的期望价值数据。

第三节　概率论在油气勘探风险决策中的应用

一、决策理论基础

决策是为解决当前或未来可能发生的问题选择最佳方案的一种过程，它贯穿于管理的全过程，因而，可以说决策是管理的核心。传统的决策乃是经验的决策，它仅依赖于决策人员个人的经验和勇气。近30年来，由于电子计算机技术的迅速发展，使得人们能够运用概率论及数理统计的原理对需决策的问题进行定量化分析并从中优选出最佳方案。目前，决策科学已成为管理科学和应用统计学的一个主要组成部分。

1．决策分类

按照对自然状态的认识程度，待决策问题可分为三类。

1)确定型决策

在决策时有完全充分的情报或信息可以把握未来的客观环境，即未来的自然状态是完全肯定的这样一类决策，这类决策常采用线性规划、动态规划和网络规划等最优化技术。

2)非确定型决策

在决策时，只知道未来可能存在的几种自然状态，但这些自然状态以多大的可能性出现则一无所知。在进行非确定型决策时，需要知道损益矩阵(表 1－5)。非确定型决策方法包括最大收益值法(乐观法)、最大最小收益值法(悲观法，Wald 决策准则)、最小最大后悔值法(Regret Value)和等可能决策准则(Laplace 准则)。

表 1－5　损益矩阵

损益值 \ 自然状态 / 备选方案	S_1	S_2	…	S_n
A_1	V_{11}	V_{12}	…	V_{1n}
A_2	V_{21}	V_{22}	…	V_{2n}
·	…	…	…	…
A_m	V_{m1}	V_{m2}	…	V_{mn}

对于同样一个待决策问题，使用不同的方法一般将导致不同的最佳决策方案。由于对自然状态出现的可能性大小一无所知，所以，很难制定一个评判方法好坏的客观标准。因此对于非确定型决策问题，很大程度上仍然依赖于决策者持有的心理状况。为了使决策更加合理，一般应对自然状态出现的概率加以预测，进而使用较为客观的尺度对决策方案进行取舍，而这就是接着要谈的风险型决策问题。

3)风险型决策

在通常的情况下，人们虽然不能完全确定未来的自然状态是什么，但总可以通过各种手段预测出未来自然状态出现的概率。风险型决策亦称统计型决策或随机型决策，进行这类决策，除已知损益矩阵外，还必须掌握各自然状态发生的概率(表 1－6)。风险型决策可分为最大可能法和期望值决策法。尽管依据概率的决策必然会有风险，但是由于遵从了统计规律，所以这种方法比凭直观想象进行决策要合理和有效得多，是一种较好的常用的决策方法。

表 1－6　已知各自然状态出现概率条件下的损益矩阵

损益值 \ 自然状态 / 备选方案	S_1, P_1	S_2, P_2	…	S_n, P_n
A_1	V_{11}	V_{12}	…	V_{1n}
A_2	V_{21}	V_{22}	…	V_{2n}
·	…	…	…	…
A_m	V_{m1}	V_{m2}	…	V_{mn}

(1)最大可能决策法。

最大可能法将风险型决策问题当作确定型问题来处理，鉴于已知各自然状态发生的概率，最大可能法要求将发生概率最大的自然状态作为必然出现的自然状态来对待，而将其它发生概率较小的自然状态视为不可能事件。

(2)期望值决策法。

期望值决策法把每个备选方案都看成是随机变量，其取值就是该方案对应于各自然状态的损益值，即损益矩阵中同一行的各元素(表 1－5)。期望值决策法在具体解题时可采用多种分析问题的形式，如决策表、决策树等。

2.灵敏度分析

在一定范围内变动各自然状态出现的概率及相应各方案的损益的大小而对最佳决策方案的稳定性进行讨论。在此,需引入转折概率的概念,所谓转折概率是指每一备选方案在不同自然状态下的损益值都不发生变化的前提下,使各个备选方案具有相同损益期望值的自然状态出现的概率。灵敏度分析指出了最佳方案的适应性和稳定性。

3.效用理论

在风险型决策中,同量的损益值因决策的风险不同会给同一决策人员带来不同的满足程度,效用曲线有3种类型(图1-5)。

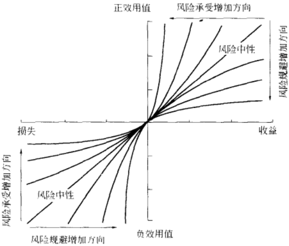


图1-5 效用曲线

图1-5中的直线反映一种严格按损益期望值决策的人对待风险的态度。向左上方凸出的曲线反映一种较为保守的决策人员对待风险的态度,从本质上来说,这类人员对损失比较敏感,对利益的反应比较迟缓,对风险持一种厌恶态度。向右下方凸出的曲线反映一种富于进取精神的决策人员对待风险的态度,从本质上来说,这类人员对利益比较敏感,对损失不太介意,是一种敢冒风险的人物。

对大量统计资料的分析表明,大多数的决策人员属于保守型的,敢冒风险和严格按期望值大小进行决策的人仅占少数。

二、基于期望货币值的决策方法

基于期望货币值决策方法的原理,是计算出每个方案的期望货币价值并选择期望货币价值最大的方案,用决策树可使这一过程清晰明了,决策树是概率树在风险决策中的应用,下面用一个例子说明该方法的应用。

【例1-7】 某油公司就某一勘探区块有如下决策选择:(1)立即卖掉区块勘探开发权,获得125k \$;(2)一年后卖掉区块勘探开发权,若油价下降,获100k \$(折现后),且油价下降的概率为0.9,若油价升高,获400k \$(折现后),且油价升高的概率为0.1;(3)投资200k \$钻探,钻探结果出现干井、低产井和高产井的概率值和纯收益(折现后)列于表1-7中。作为决策者,

在三个方案中,您如何选择?

表 1-7　例 1-7 中的有关数据

井类型	概率	纯收益
干井	0.50	-200k
低产	0.40	200k
高产	0.10	1300k

解:(1)计算各方案的期望货币值:

(a)立即卖掉区块勘探开发权 $EMV = 125k\$$;

(b)一年后卖掉其勘探开发权 $EMV = 0.1 \times 400k\$ + 0.9 \times 100k\$ = 130k\$$;

(c)投资 200k$,钻探 $EMV = 0.5 \times (-200k\$) + 0.4 \times 200k\$ + 0.1 \times 1300k\$ = 110k\$$。

(2)各方案期望货币值比较:

通过比较,一年后卖掉区块勘探开发权具有最大的期望货币值,故选择一年后卖掉区块勘探开发权。

上述决策过程可利用决策树(图 1-6)清晰明了地图示出来。

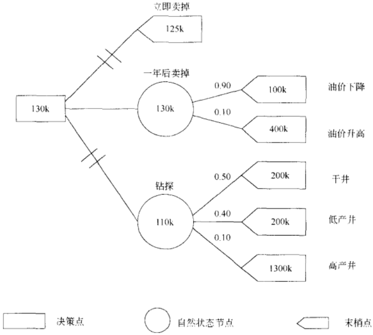


图 1-6　决策树应用实例分析

三、基于期望效用值的决策方法

上面根据最大期望货币值进行的决策,反映的是图 1-5 中的直线所对应的对风险具有中性态度的决策者所做的决策。但不同的决策者对待风险的态度是不同的,不同的人有不同的效用曲线,根据期望效用值决策的结果也是不同的(图 1-7)。

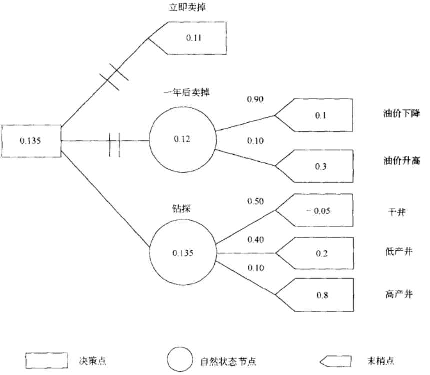


图 1－7　基于期望效用值的决策

在图 1－7 中，末梢点上将期望货币值用期望货币值带给某个决策者的期望效用值来取代，最终得到的期望效用值表明，选择“钻探”给该决策者带来最大的期望效用值（即最大的满足度），故按期望效用值决策方法，该决策者会选择“钻探”而非“一年后卖掉”。

四、贝叶斯(Bayes)决策方法

当一些新的信息获得后，决策会发生变化，如何把新的信息“融入”决策过程，是贝叶斯决策方法要解决的问题，下面，我们用实例来说明贝叶斯决策方法。

【例 1－8】　在例 1－7 中，假定还存在第四个方案，即在钻井之前部署地震工作量以进一步落实地质构造，其费用为 10k $，地震工作量的结果有三种可能，一是无构造，二是构造开启，三是构造封闭，且根据以往勘探的经验，不同类型井对应的地质构造情况列于表 1－8。作为决策者，在四个方案中，您如何选择？

表 1－8　例 1－8 中不同井类型对应的地质构造情况

	无构造	构造开启	构造封闭
干井	0.70	0.20	0.10
低产	0.30	0.40	0.30
高产	0.10	0.30	0.60

解：(1)首先，在做地震工作量之前，需要知道“无构造”、“构造开启”和“构造封闭”出现的

概率,可根据表1-7和表1-8提供的信息,用图1-8的概率树计算出这些概率值:

$$P(\text{无构造})=0.48, P(\text{构造开启})=0.29, P(\text{构造封闭})=0.23$$

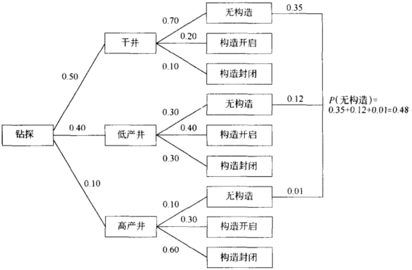


图1-8 概率树的应用

(2)其次,需要知道在出现"无构造"、"构造开启"和"构造封闭"不同情况的条件下,出现"干井"、"低产井"和"高产井"的概率,这些概率值可用贝叶斯公式(1-8)求得,如在"无构造"事件出现的情况下,"干井"事件出现的概率为:

$$P(\text{干井}/\text{无构造})=P(\text{干井和无构造同时出现})/P(\text{无构造})=0.35/0.48=0.73$$

求取的概率值表示在图1-9的逆概率树中。

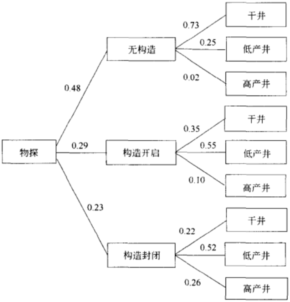


图1-9 逆概率树的应用

(3)根据以上信息,就可绘制图 1－10 所示的决策树,该图表明,选择"钻井之前部署地震工作量"将得到最大的期望货币值。

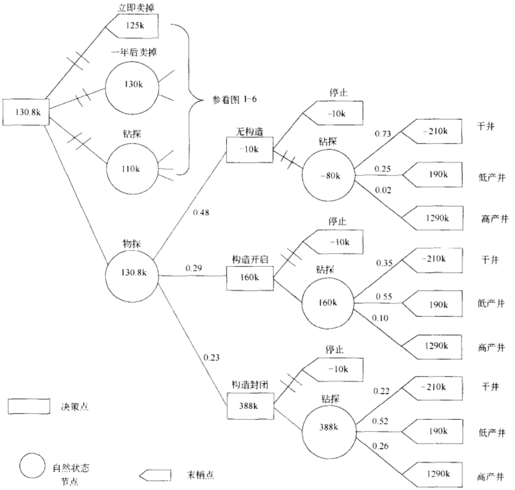


图 1－10 贝叶斯决策树

◆ 思 考 和 练 习 题 ◆

1.如何理解和区分两个事件之间的"不相容性"和"独立性"? 在两个事件相容的情况下,概率的加法公式应如何写?

2.如何理解条件概率公式、概率的乘法公式和全概率公式,并从这些公式出发,推导出贝叶斯概率公式。

3.某油层中有三种沉积微相,已知其中第一、第二和第三种沉积微相各占总量的 50%、25%和 25%,又知第一、第二和第三种沉积微相高渗透($>100\times10^{-3}\mu m^2$)岩心出现的概率分别为 1%、2%和 4%,根据例 1－6 的计算,当已知取出的一个样品为高渗透率样品时,这个样品

属于第一、第二和第三种沉积微相的可能性分别为 25%、25% 和 50%，问哪些是先验概率？哪些是后验概率？

4.就图 1－10 中的决策，若考虑您个人的效用曲线（自己为自己构筑此曲线），您的最终决策选择是什么？

5.某勘探项目组在某一区块设计了 4 口井的勘探方案，根据经济评价情况，4 口井中至少有两口井发现油气，这一方案才是经济有效的，假定钻第一口井的成功率为 0.1，并假定在每一成功井之后的探井成功的概率增加一倍，问这个勘探方案成功的概率有多大？

6.有 x 型汽车，出厂时次品占 30%。对合格的车，某试车手在试车后，有 80% 的可能认为是合格车，20% 的可能认为是次品车；对不合格的车，有 90% 的可能认为是次品车，10% 的可能认为是合格车。问(1)直接购买得到次品车的可能性为多大？(2)经试车手试车后认为合格再购买的情况下，得到次品车的可能性为多大？

第二章 随机变量和统计推断

本章学习目的和要求：本章介绍随机变量、随机变量的数字特征、随机变量的分布函数、参数估计和假设检验及其在油气藏地质参数研究中的应用，要求学生能够用这些基本工具解决在工作中遇到的相关问题。学习本章内容，需要具备第一章中概率论的基本知识。

第一节 随机变量及其数字特征

一、随机变量

1.随机变量的定义

设随机试验 E 的样本空间为 $\Omega=\{\omega\}$，$Z=Z(\omega)$是定义在 Ω 上的单实值函数，而且对任一实数 z，事件$\{Z\leqslant z\}$都有确定的概率，则称 Z 是一个随机变量，记为 $Z=Z(\omega)$，简言之，随机变量就是具有一定概率分布的变量。

从数学的角度来看，随机变量与样本空间实际上是一个对应关系，即将样本空间中每一个可能的结果与一个实数相对应(图 2－1)。

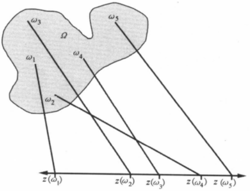


图 2－1 随机变量与样本的对应关系

【例 2－1】 设一个盒中有 5 件产品，其中 3 件为正品(标号为 1、2 和 3)，2 件为次品(标号为 4 和 5)，从中取出 3 件，观察取出的 3 件产品中含正品的件数，含正品的件数即为一随机变量 Z(表 2－1)。

在本例中，取出的三件产品中含正品的件数与样本空间中的每个样本一一对应，且对任一实数 z，事件$\{Z\leqslant z\}$都有确定的概率，如 $P\{Z\leqslant 1\}=0.3$，符合随机变量的定义，故 Z 是一个随机变量。

表 2-1　随机试验的样本空间及其对应的随机变量

ω	Z(正品件数)	ω	Z(正品件数)
(1,2,3)	3	(1,4,5)	1
(1,2,4)	2	(2,3,4)	2
(1,2,5)	2	(2,3,5)	2
(1,3,4)	2	(2,4,5)	1
(1,3,5)	2	(3,4,5)	1

2.随机变量的分布函数和密度函数

要研究随机变量 Z,首先要掌握 Z 在各个范围取值的概率,为此,我们引入分布函数的概念,随机变量 Z 的分布函数定义为

$$F(z) = P\{Z \leqslant z\} \qquad -\infty < z < +\infty \tag{2-1}$$

即分布函数是事件$\{Z \leqslant z\}$发生的概率 P。

$F(z)$具有如下基本特征(图 2-2):

(1)$0 \leqslant F(z) \leqslant 1$;

(2) $\lim\limits_{z \to -\infty} F(z) = 0$;

(3) $\lim\limits_{z \to \infty} F(z) = 1$;

(4)$F(z)$关于自变量 z 单调不减。

注意 $F(z)$实质上是累积分布函数,简写为 *cdf*。

在地质应用研究中,常常对变量值大于某一给定值发生的概率感兴趣,即用到"大于"概率分布函数,记为 $AF(z)$

$$AF(z) = P\{Z > z\} = 1 - F(z) \tag{2-2}$$

分布函数 $F(z)$的导数,为随机变量 Z 的概率密度函数,简写为 *pdf*(图 2-3),即

$$f(z) = F'(z) \qquad -\infty < z < +\infty \tag{2-3}$$

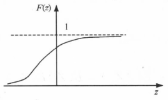


图 2-2　随机变量的(累积)分布函数

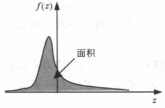


图 2-3　随机变量的概率密度函数

概率密度函数 $f(z)$具有以下特点:

(1) $\lim\limits_{z \to -\infty} f(z) = 0$;

(2) $\lim\limits_{z \to \infty} f(z) = 0$;

(3)$f(z)$曲线与 z 轴之间的面积为 1;

(4)某一 z 值左侧的面积等于(累积)分布函数 $F(z)$。

3.随机变量的数字特征

随机变量的数字特征包括期望值、方差和 k 阶矩。

随机变量 Z 的数学期望值是 Z 所取的全部可能值的加权平均值,是刻画随机变量 Z 取值平均大小的一个数字特征

$$E(Z)=\mu=\int_{-\infty}^{+\infty}z\mathrm{d}F(z)=\int_{-\infty}^{+\infty}zf(z)\mathrm{d}z \tag{2-4}$$

随机变量 Z 的方差是刻画随机变量 Z 在它的数学期望周围取值分散或集中程度的一个数,是随机变量 Z 的另一个数字特征

$$Var(Z)=E[Z-E(Z)]^2=\int_{-\infty}^{+\infty}(z-\mu)^2\mathrm{d}F(z)=\int_{-\infty}^{+\infty}(z-\mu)^2f(z)\mathrm{d}z=\sigma^2 \tag{2-5}$$

随机变量 Z 的方差,也可展开成以下形式

$$Var(Z)=E[Z-E(Z)]^2=E(Z^2)-E(Z)^2=\sigma^2 \tag{2-6}$$

随机变量 Z 的 k 阶原点矩由下式定义

$$E(Z^k)=\int_{-\infty}^{+\infty}z^k\mathrm{d}F(z)=\int_{-\infty}^{+\infty}z^kf(z)\mathrm{d}z \tag{2-7}$$

式中,k 为正整数。显然,随机变量 Z 的 1 阶原点矩即为随机变量 Z 的期望值。

随机变量 Z 的 k 阶中心矩由下式定义

$$E[Z-E(Z)]^k=\int_{-\infty}^{+\infty}[z-E(Z)]^k\mathrm{d}F(z)=\int_{-\infty}^{+\infty}[z-E(Z)]^kf(z)\mathrm{d}z \tag{2-8}$$

式中,k 为正整数。显然,随机变量 Z 的 2 阶中心矩即为随机变量 Z 的方差。

二、n 维随机向量

1.n 维随机向量的定义和协方差函数

在实际问题中,常常需要研究多个随机变量之间的关系,这就要借助于随机向量。n 维随机向量是由 n 个随机变量组成的变量组,记为$\{Z_1,Z_2,\cdots,Z_n\}$。

n 维联合分布反映变量之间的相互关系

$$F(z_1,z_2,\cdots,z_n)=P\{Z_1\leqslant z_1,Z_2\leqslant z_2,\cdots,Z_n\leqslant z_n\} \tag{2-9a}$$

n 维联合分布的概率密度函数定义为

$$f(z_1,z_2,\cdots,z_n)=F'(z_1,z_2,\cdots,z_n) \tag{2-9b}$$

任意两个变量之间的协方差函数定义为

$$\begin{aligned}Cov(Z_j,Z_{j'})&=E\{[Z_j-E(Z_j)][Z_{j'}-E(Z_{j'})]\}=E(Z_jZ_{j'})-E(Z_j)E(Z_{j'})\\&=\sigma_{z_jz_{j'}}\end{aligned} \tag{2-10}$$

协方差函数反映两个变量之间的相关性,$\mathrm{Cov}(Z_j,Z_{j'})>0$,变量 $Z_j,Z_{j'}(j\neq j')$正相关,$\mathrm{Cov}(Z_j,Z_{j'})<0$;变量 $Z_j,Z_{j'}(j\neq j')$负相关;$\mathrm{Cov}(Z_j,Z_{j'})=0$,变量 $Z_j,Z_{j'}(j\neq j')$不相关。

任意两个随机变量之间的相关系数定义为

$$\gamma_{Z_jZ_{j'}}=\frac{Cov(Z_j,Z_{j'})}{\sqrt{Var(Z_j)Var(Z_{j'})}}\quad(j\neq j') \tag{2-11}$$

相关系数也反映两个变量之间的相关性,且消除了量纲,$-1\leqslant\gamma_{Z_jZ_{j'}}\leqslant1$,$|\gamma_{Z_jZ_{j'}}|$越大,相关性越强,反之亦然。

【例 2-2】 在例 2-1 中,同时考虑所取 3 件产品中的正品件数和所取 3 件产品的编号之和,即构成 2 维随机向量$\{Z_1, Z_2\}$(表 2-2)。试求 Z_1, Z_2 的联合分布、协方差及相关系数,判别这两个变量的相关性。

表 2-2 随机试验的样本空间及其对应的二维随机向量$\{Z_1, Z_2\}$

ω	Z_1 (正品件数)	Z_2 (编号之和)	ω	Z_1 (正品件数)	Z_2 (编号之和)
(1,2,3)	3	6	(1,4,5)	1	10
(1,2,4)	2	7	(2,3,4)	2	9
(1,2,5)	2	8	(2,3,5)	2	10
(1,3,4)	2	8	(2,4,5)	1	11
(1,3,5)	2	9	(3,4,5)	1	12

解:可计算概率值 $P\{Z_1 = l, Z_2 = k\}\begin{pmatrix} l = 1,2,3 \\ k = 6,7,\cdots,12 \end{pmatrix}$,得$(Z_1, Z_2)$的联合概率密度以及边缘概率密度,列于表 2-3 中。

表 2-3 Z_1 和 Z_2 两个随机变量的联合概率密度及边缘概率密度

Z_1 \ f \ Z_2	6	7	8	9	10	11	12	$P\{Z_1 = l\}$
1	0	0	0	0	1/10	1/10	1/10	3/10
2	0	1/10	2/10	2/10	1/10	0	0	6/10
3	1/10	0	0	0	0	0	0	1/10
$P\{Z_2 = k\}$	1/10	1/10	2/10	2/10	2/10	1/10	1/10	1

要计算协方差函数,需先计算随机变量 Z_1 和 Z_2 的数学期望和方差

$$E(Z_1) = \sum_{l=1}^{3} lP\{Z_1 = l\} = 1.8 \qquad E(Z_2) = \sum_{k=6}^{12} kP\{Z_2 = k\} = 9$$

$$\mathrm{Var}(Z_1) = E[Z_1 - E(Z_1)]^2 = 0.36$$

$$\mathrm{Var}(Z_2) = E[Z_2 - E(Z_2)]^2 = 3$$

则协方差函数为:$\mathrm{Cov}(Z_1, Z_2) = E\{[Z_1 - E(Z_1)][Z_2 - E(Z_2)]\} = -0.9$

相关函数为:$\gamma_{Z_1 Z_2} = \dfrac{\mathrm{Cov}(Z_1, Z_2)}{\sqrt{\mathrm{Var}(Z_1)\mathrm{Var}(Z_2)}} = \dfrac{-0.9}{\sqrt{0.36 \times 3}} = -0.274$

从协方差函数取负值来判断,随机变量 Z_1 和 Z_2 之间呈负相关关系,从相关函数的取值来判断,随机变量 Z_1 和 Z_2 之间呈较弱的负相关关系。

2. n 维随机向量的物理含义

对 n 维随机向量可以有 3 种理解:(1) 不同属性的变量在同一点上;(2) 同一属性的变量在不同点上;(3) 不同属性的变量在不同点上。

若按第一种含义去理解,不涉及空间问题,只涉及多个变量之间的关系问题,属多元统计学研究的范畴,多元统计学方法将在第三章后半部分和第四章介绍。

若按第二种理解,则 n 维随机向量就是一个随机函数,记为 $Z(x)$,$Z(x)$可称作空间或时间上相关的一组随机变量,x 为空间位置,也可理解为所有空间的集合。随机函数在一维时间

轴上，称随机过程，在二维或三维空间轴上称随机场，若把时间和空间问题结合起来，就构成一个时空模型。电场、磁场、重力场和温度场等都是空间点的函数，且它们也都有随机性的变化，故都可看成随机场。对它们每次随机观测的结果，都得到一个确定的空间点函数(或实现)，而当我们对空间固定的某点 x_0 来考察它们时，则得到一个随机变量 $Z(x_0)$。随机函数及其在地质量变研究中的应用，将在第五章至第七章讨论。

若按第三种理解，则 n 维随机向量为多元区域化变量和多元地质统计学研究的对象，多元地质统计学超出本教材讨论的范围，有兴趣的读者可参考有关文献。

第二节　常见的概率分布函数

一、正态分布

正态分布的概率密度函数为钟形(对称)曲线。正态分布有两个参数，即随机变量的期望值 μ 和方差σ^2。若随机变量 Z 服从正态分布，可记为 $Z \sim N(\mu,\sigma^2)$。期望值 μ 为曲线的最高点对应的横坐标值。方差 σ^2 反映随机变量的离散程度，σ^2 越小，曲线就越瘦，随机变量的分布就越集中；σ^2 越大，随机变量的分布越离散，概率密度函数曲线就越胖。图 2－4 中的曲线为 $\mu=0$ 和 σ^2 分别取不同值的三个正态分布的概率密度曲线($\sigma_1^2<\sigma_2^2<\sigma_3^2$)。

在油气资源评价中，正态分布随机变量起了十分重要的作用，孔隙度、有效油层厚度等常呈正态分布形式。

正态分布的密度函数为

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right] \tag{2-12}$$

式中　μ——随机变量 Z 的期望值；

σ——随机变量 Z 的均方差。

正态分布的累积分布函数为

$$F(z) = \int_{-\infty}^{z}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]\mathrm{d}t \tag{2-13}$$

式(2－12)中，当 $\mu=0,\sigma^2=1$ 时，得到标准正态分布[记为 $Z\sim N(0,1)$]的密度函数

$$f(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \tag{2-14}$$

标准正态分布的(累积)分布函数为

$$F(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2}\right]\mathrm{d}t \tag{2-15}$$

从标准正态分布表(附表 1)中读出的对应 z 值的数值，实际上是 $F(z)$值或者说是 z 值左侧概率密度曲线的面积(图 2－5)，即事件$\{Z\leqslant z\}$发生的概率值。

借助于标准正态分布表，我们可以非常方便地解决有关的概率求解问题，对任一服从正分布态的变量 $Z\sim N(\mu,\sigma^2)$，可通过以下变换

$$z' = \frac{z-\mu}{\sigma} \tag{2-16}$$

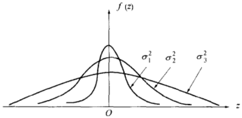


图 2-4 正态分布曲线

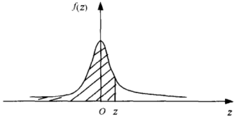


图 2-5 标准正态分布表中的数值对应于 z 值左侧的面积

变为标准正态分布的变量 $Z' \sim N(0,1)$，也就可以借助于标准正态分布表非常方便地解决有关的概率求解问题了。

【例 2-3】 某油层有大量的岩心孔隙度资料，根据对这些资料的分析发现，孔隙度值服从正态分布，数学期望 $\mu = 10$，方差 $\sigma^2 = 9$。现要打一口新井，问(1)获得该油层岩心孔隙度大于 15%的概率有多大？(2)60%的可能取到小于某一孔隙度值的岩心，这一孔隙度值是多大？

解：(1)获得该油层岩心孔隙度大于 15%的概率是事件$\{Z > 15\%\}$发生的概率，即

$$P\{Z > 15\%\} = 1 - P\{Z \leqslant 15\%\} = 1 - F(15)$$

现在来求 $F(15)$，首先用式(2-16)的变换把正态分布变量 $Z \sim N(10,9)$转化标准正态分布变量 $Z' \sim N(0,1)$，对应于 $z = 15$，得到 $z' = 1.67$，从标准正态分布表中可读出

$$F(15) = P\{Z \leqslant 15\} = P\{Z' \leqslant 1.67\} = F(1.67) = 0.9525，故有$$

$$P(Z > 15\%\} = 1 - P\{Z \leqslant 15\%\} = 1 - F(15) = 0.0475$$

(2)在标准正态分布表中读出 60%概率对应的 $z' \approx 0.25$，故对应的孔隙度值 $z = \mu + \sigma z' = 10\% + 3\% \times 0.25 = 10.75\%$。

二、对数正态分布

当某一随机变量的对数值服从正态分布时，则该变量服从对数正态分布，在油气勘探评价中，对数正态分布得到了广泛的应用，如渗透率、油层厚度、采油率、储量和油田规模等常属于这种分布。

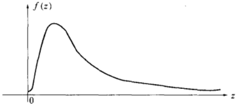


图 2-6 对数正态分布曲线

对数正态分布是一连续型概率分布，概率密度曲线呈非对称右偏(图 2-6)，概率密度为

$$f(z) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\ln z - \mu)^2}{2\sigma^2}} & (z > 0) \\ 0 & (z \leqslant 0) \end{cases} \tag{2-17}$$

式中，$\mu = E(\ln Z)$，$\sigma^2 = \mathrm{Var}(\ln Z)$。

三、三角分布

当原始数据只有最小值 a、最大值 c 和介于二者之间的最佳值 b 且不知道随机变量的分布概型时，一般采用三角分布函数。三角分布是一种三角形的连续分布(图2-7)，三角分布的概率密度为

$$f(z)=\begin{cases}\dfrac{2(z-a)}{(c-a)(b-a)} & a\leqslant z\leqslant b\\ \dfrac{2(c-z)}{(c-a)(c-b)} & b<z\leqslant c\\ 0 & \text{其他}\end{cases} \tag{2-18}$$

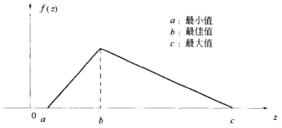


图 2-7　三角分布图

四、均匀分布

当随机变量的数据只有最小值 a 和最大值 b，一般采用最简单的均匀分布来代替随机变量的分布函数。均匀分布是描述随机变量的每一个数值在某一区间$[a,b]$内等可能发生的连续型概率分布，其概率密度为

$$f(z)=\begin{cases}\dfrac{1}{b-a} & a<z<b\\ 0 & z\leqslant a\text{ 或 }z\geqslant b\end{cases} \tag{2-19}$$

第三节　抽样及样本统计量的分布

一、总体、样本及样本统计量

1.总体(母体)

服从某一定义的所有个体的集合称为总体(母体)，在数理统计中，称随机变量或其概率分布为总体，记为 Z。总体可以是有限总体(如某年级某班所有的学生)，也可以是无限总体(如某油层的孔隙度)。

2.样本(子样)

为了研究总体的性质，乍看起来，最好是把每个个体的要研究的性质都加以测定。但这往往是不必要的，有时甚至是不可能的。我们常常从总体中抽取 n 个个体进行观测，然后就根据这 n 个观测值去推断总体的性质，由 n 个个体构成的集合称为样本(子样)，n 小于总体所包含的个体数，n 称为样本容量。样本是通过抽样获得的，抽样的目的是通过获得的样本对总体的特征进行统计推断。要使这种推断是合理的，首先就要求抽样过程是随机的，是经过精心设计的。如要研究某油层的孔隙度，就要尽量使样本的个体在空间上均匀分布，否则，得到的样本就不具有代表性。

3.样本分布函数(经验分布函数)和频率直方图

母体 Z 的分布函数 $F(z)$ 往往是未知的，如何根据来自母体 Z 的一个样本去估计 $F(z)$

呢？我们可以这么办，将具体观测到的结果（即样本观测值）$z_1, z_2, \cdots, z_i, \cdots, z_n$ 按大小顺序排列成 $z_1^* \leqslant z_2^* \leqslant \cdots \leqslant z_i^* \leqslant \cdots \leqslant z_n^*$，定义函数

$$F_n(z) = \begin{cases} 0 & z < z_1^* \\ \dfrac{i}{n} & z_i^* \leqslant z \leqslant z_{i+1}^*, (i = 1,2,\cdots,n-1) \\ 1 & z > z_n^* \end{cases} \tag{2 - 20}$$

其图形如下：

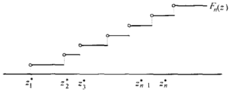


图 2 - 8　样本（累积）分布函数

显然，对给定的样本观测值 $z_1, \cdots, z_n$（从而也就给定了 $z_1^*, z_2^*, \cdots, z_n^*$）来说，$F_n(z)$作为 z 的函数，是一个不降、左连续函数，且 $F_n(-\infty) = 0, F_n(+\infty) = 1$。故 $F_n(z)$是一个概率分布函数，称其为样本分布函数（或经验分布函数）。

而对每一个给定的 z 来说，$F_n(z)$是在对 Z 的 n 次独立观测中，小于或等于 z 的观测值出现的频率，亦即事件$\{Z \leqslant z\}$出现的频率，它是一个随机变量。根据贝努里大数定律可知 $F_n(z)$ 依概率收敛到事件$\{Z \leqslant z\}$的概率，即 $F_n(z) \xrightarrow{P} F(z)(n \to \infty)$。因此，当 n 充分大时，样本函数 $F_n(z)$是母体分布函数 $F(z)$的良好近似。这正是人们常用 $F_n(z)$去估计 $F(z)$的原因。图 2 - 9 是根据 100 个灯泡的使用寿命所画出的样本分布函数 $F_{100}(z)$及其相应的母体分布函数 $F(z)$。其中阶梯形函数为 $F_{100}(z)$。这里 $n = 100$ 比较大，由图形可以大致看出，$F_{100}(z)$是 $F(z)$的良好近似。

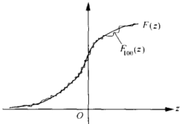


图 2 - 9　根据 100 个灯泡的使用寿命所画出的样本分布函数 $F_{100}(z)$
及其相应的母体分布函数 $F(z)$

构造样本分布函数，无论是连续型随机变量还是离散型随机变量都是可以的。但对连续型母体 Z 来说，人们常用频率直方图去给出母体 Z 的分布密度$f(z)$的良好近似。

如前，将对 Z 的 n 次独立观测值 $z_1, z_2, \cdots, z_n$ 依大小顺序排成 $z_1^* \leqslant z_2^* \leqslant \cdots \leqslant z_n^*$，取略小于 z_1^* 的数 a 及略大于 z_n^* 的数 b，并在区间$(a, b]$中插入一些分点 $a = t_0 < t_1 < \cdots < t_{m-1} < t_m = b$，将区间$(a, b]$划分成 m 个左开右闭的小区间。实践中通常取分点为等分点，且使落入每

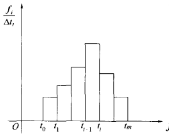


图 2-10　频率直方图

个小区间内的点不少于5个(否则,可将相邻的若干小区间加以合并,使每个小区间内的数据不少于5个。当然,这样分点就不一定是等分点了,但这也无妨)。我们以 ν_i 表示落入第 i 个小区间$(t_{i-1}, t_i]$内的观测值的个数——称为频数,则 $\nu_i/n = f_i$ 就是观测数据中落入第 i 个小区间内的频率。在直角坐标系内,以各个小区间为底,画一竖着的长方形,使第 i 个长方形的面积为 f_i,通常称这样的图形为频率直方图,简称为直方图(见图2-10)。

由大数定律知,f_i 作为母体 Z 的观测值落入第 i 个小区间$(t_{i-1}, t_i]$内的频率,应近似于(在依概率收敛的意义下)Z 落入第 i 个小区间内的概率,即应有:

$$f_i \approx P(t_{i-1} < Z \leqslant t_i) = \int_{t_{i-1}}^{t_i} f(z)\mathrm{d}z \tag{2-21}$$

若 Z 的概率密度$f(z)$在$(t_{i-1}, t_i]$内连续,则

$$\int_{t_{i-1}}^{t_i} f(z)\mathrm{d}z \approx f(z)\Delta t_i \tag{2-22}$$

其中,z 是$(t_{i-1}, t_i]$内的任意一点,$\Delta t_i = t_i - t_{i-1}$。

从而 $f_i \approx f(z)\Delta t_i$,即 $f_i/\Delta t_i \approx f(z)$。由此可见,频率直方图的上部轮廓线——折线,确是密度函数 $f(z)$的良好近似。

4.样本统计量

样本统计量是不含未知参数的样本的函数,是由样本观测值计算得到的,如样本均值

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{2-23}$$

和样本方差

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{Z})^2 \tag{2-24}$$

都是样本统计量,还有样本的 k 阶原点矩和 k 阶中心矩也是样本的统计量,这些样本的统计量统称为样本的数字特征。

二、样本统计量的分布和中心极限定理

要根据样本对总体的性质进行统计推断,还必须了解样本统计量的分布特征,这里主要讨论样本均值的分布特征。

假如从总体中随机抽取若干个容量为 n 的样本,对每个样本都可以计算均值,这些均值的分布即为样本均值的分布,下面,我们通过一个简单的算例引出一个重要的定理——中心极限定理。

【例 2-4】 设某一随机变量 Z(总体)包含5个个体(样本点),其变量值、期望值和方差列于表2-4中,随机抽取容量为2的样本,考察样本均值的分布。

表 2-4　随机变量 Z(总体)的取值、期望值和方差

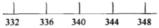
序号	Z 的取值	Z 的期望值	Z 的方差
1	332		
2	336		
3	340	$\mu = E(Z) = \frac{1}{5}\sum_{i=1}^{5} z_i = 340$	$\text{Var}(Z) = \sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(z_i - 340)^2 = 32$
4	344		
5	348		

解:随机抽取容量为 2 的样本(放回抽样),可能抽到的样本点的组合、变量值、样本均值及样本统计量列于表 2-5 中。

表 2-5　样本点的组合、变量值、样本均值及样本均值统计量

样本序号	样本点组合	变量值组合	样本均值($\bar{Z}$)	样本均值统计量
1	(1,1)	(332,332)	332	
2	(1,2)	(332,336)	334	
3	(1,3)	(332,340)	336	
4	(1,4)	(332,344)	338	
5	(1,5)	(332,348)	340	
6	(2,1)	(336,332)	334	
7	(2,2)	(336,336)	336	
8	(2,3)	(336,340)	338	
9	(2,4)	(336,344)	340	
10	(2,5)	(336,348)	342	样本均值的期望值
11	(3,1)	(340,332)	336	$E(\bar{Z}) = \frac{1}{25}\sum_{i=1}^{25}\bar{Z}_i = 340$
12	(3,2)	(340,336)	338	样本均值的方差
13	(3,3)	(340,340)	340	$Var(\bar{Z}) = \frac{1}{25}\sum_{i=1}^{25}[\bar{Z}_i - E(\bar{Z})]$
14	(3,4)	(340,344)	342	$= 16 = \frac{32}{2}$
15	(3,5)	(340,348)	344	
16	(4,1)	(344,332)	338	
17	(4,2)	(344,336)	340	
18	(4,3)	(344,340)	342	
19	(4,4)	(344,344)	344	
20	(4,5)	(344,348)	346	
21	(5,1)	(348,332)	340	
22	(5,2)	(348,336)	342	
23	(5,3)	(348,340)	344	
24	(5,4)	(348,344)	346	
25	(5,5)	(348,348)	348	

对比表 2-5 和表 2-4 我们发现,样本均值的期望值正好等于变量(总体)的期望值,样本均值的方差正好等于变量(总体)的方差除以样本容量。对比样本均值的概率密度函数和变量



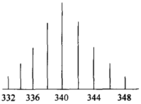


图 2-11 随机变量(总体)和样本均值概率密度函数对比

的概率密度函数,见图 2-11(a)、(b),我们发现样本均值的概率密度函数趋于对称。这些现象的出现不是巧合,而是背后的定理发挥"操控"作用的结果,这一定理就是著名的中心极限定理。

中心极限定理可表述为:从期望值为 μ、方差为 σ^2 的随机变量(总体)中随机地抽取容量为 n 的样本,该样本的均值 $\bar{Z}$ 为一随机变量,该随机变量趋向于服从期望值为 μ、方差为$\frac{\sigma^2}{n}$的正态分布,即

$$
\begin{aligned}
E(\bar{Z}) &= \mu \\
Var(\bar{Z}) &= \frac{\sigma^2}{n}
\end{aligned}
\tag{2-25}
$$

且这种趋势随 n 的增大而愈加明显。

根据具体的应用条件,可分为 3 种情况:

(1)总体服从正态分布,且 σ^2 已知

此时,$\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}}$ 服从标准正态分布。

(2)总体服从任意分布,σ^2 未知,$n \geqslant 30$

此时,$\frac{\bar{Z}-\mu}{s/\sqrt{n}}$ 近似服从标准正态分布,其中,$s=\sqrt{\frac{\sum_{i=1}^{n}(z_i-\bar{Z})^2}{n-1}}$为样本均方差。

(3)总体服从正态分布,σ^2 未知,$n<30$

此时,中心极限定理已不适应,$\frac{\bar{Z}-\mu}{s/\sqrt{n}}$ 服从自由度为$(n-1)$的 Student 分布(又称 t 分布,见附表 2),其中,$s=\sqrt{\frac{\sum_{i=1}^{n}(z_i-\bar{Z})^2}{n-1}}$为样本均方差。

【例 2-5】 某油层 100 块岩心孔隙度(总体)的统计资料表明,岩心孔隙度(总体)服从正态分布,数学期望 $\mu=12$,方差 $\sigma^2=16$。从这 100 个岩心孔隙度(总体)数据中随机抽取 25 个岩心孔隙度数据作为一个样本,问该样本均值有多大的概率落在[10,15]这一区间内?

解:根据中心极限定理,样本均值 $\bar{Z}$ 服从正态分布,$\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}}$ 服从标准正态分布,由 $\bar{Z}=10$ 和 $\bar{Z}=15$ 可计算出 $\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}}$ 的值分别为 -2.5 和 3.75,则标准正态变量 $\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}}$ 落在$[-2.5, 3.75]$区间内的概率可由标准正态分布的分布表查得(附表 1)

$$P\left\{-2.5 \leqslant \frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant 3.75\right\} = P\left\{\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant 3.75\right\} - P\left\{\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant -2.5\right\}$$
$$= P\left\{\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant 3.75\right\} - \left(1 - P\left\{\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant 2.5\right\}\right)$$
$$= P\left\{\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant 3.75\right\} + P\left\{\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant 2.5\right\} - 1$$
$$= 0.99991 + 0.99379 - 1$$
$$= 0.9937$$

以上计算过程可参考图 2－12。

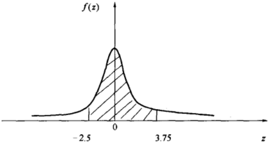


图 2－12　从标准正态分布表读取概率值示意图

第四节　统 计 推 断

统计推断是根据样本统计量及其分布特征对总体进行研究的方法，包括参数估计、假设检验和方差分析等，本节主要讨论参数估计和假设检验方法，方差分析在讨论回归分析和趋势分析具体应用时在相应的内容中讨论。

一、参数估计

1.点估计

当总体的某个参数是根据样本观测值用单一的数值来估计时，称为点估计。如式(1－23)是总体期望值 μ 的无偏估计，式(1－24)是总体方差 σ^2 的无偏估计。点估计比较简单，我们不多讨论。

2.区间估计

区间估计给出在一定置信水平 $1-\alpha$(α 为置信度)下总体某个参数出现的区间范围，如对总体期望值 μ 有如下区间估计

$$P\{\bar{Z}-K \leqslant \mu \leqslant \bar{Z}+K\} = 1-\alpha \tag{2-26}$$

要求得式(2－26)中的 K，我们先讨论 $\frac{\bar{Z}-\mu}{\sigma/\sqrt{n}}$ 服从标准正态分布的情形，从标准正态分布表(附表 1)中可读出式

$$P\left\{-z_{\alpha/2} \leqslant \frac{\bar{Z}-\mu}{\sigma/\sqrt{n}} \leqslant z_{\alpha/2}\right\} = 1-\alpha \tag{2-27}$$

中的 $-z_{\alpha/2}$ 和 $z_{\alpha/2}$(图 2－13)。

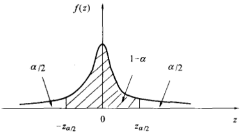


图 2－13　在 $1-\alpha$ 置信水平下从标准正态分布表中读取 $-z_{\alpha/2}$ 和 $z_{\alpha/2}$ 示意图

式(2－27)经过变换后可得

$$P\left\{\bar{Z}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\leqslant\mu\leqslant\bar{Z}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right\}=1-\alpha \tag{2-28}$$

对比式(2－26)和式(2－28),可得 $K=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$,于是,在$(1-\alpha)$置信水平下总体期望值 μ 的区间估计为

$$\bar{Z}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\leqslant\mu\leqslant\bar{Z}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2} \tag{2-29}$$

当 $\frac{\bar{Z}-\mu}{s/\sqrt{n}}$ (其中,$s=\sqrt{\frac{\sum_{i=1}^{n}(z_i-\bar{Z})^2}{n-1}}$)近似服从标准正态分布时,在$(1-\alpha)$置信水平下总体期望值 μ 的区间估计为

$$\bar{Z}-z_{\alpha/2}\frac{s}{\sqrt{n}}\leqslant\mu\leqslant\bar{Z}+z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{2-30}$$

当 $\frac{\bar{Z}-\mu}{s/\sqrt{n}}$ (其中,$s=\sqrt{\frac{\sum_{i=1}^{n}(z_i-\bar{Z})^2}{n-1}}$)服从自由度为$(n-1)$的 Student 分布时,在$(1-\alpha)$置信水平下总体期望值 μ 的区间估计为

$$\bar{Z}-t_{\alpha/2(n-1)}\frac{s}{\sqrt{n}}\leqslant\mu\leqslant\bar{Z}+t_{\alpha/2(n-1)}\frac{s}{\sqrt{n}} \tag{2-31}$$

式中的 $-t_{\alpha/2}(n-1)$ 和 $t_{\alpha/2}(n-1)$ 可从 Student 分布表(附表 2)中读取。

【例 2－6】　在某油层中取 36 块岩心作为样本,测得其孔隙度,得到样本均值 $\bar{Z}=10$,样本方差 $s^2=\frac{\sum_{i=1}^{n}(z_i-\bar{Z})^2}{n-1}=9$,问在 95%的置信水平下,该油层孔隙度(总体)的期望值落在哪个区间内?

解:根据中心极限定理,$\frac{\bar{Z}-\mu}{s/\sqrt{n}}$ 服从标准正态分布,故该油层孔隙度(总体)的期望值的置信区间可用公式 $\bar{Z}-z_{\alpha/2}\frac{s}{\sqrt{n}}\leqslant\mu\leqslant\bar{Z}+z_{\alpha/2}\frac{s}{\sqrt{n}}$ 求取,查标准正态分布表得 $Z_{\alpha/2}=Z_{0.025}=1.96$,代入题中给定的样本均值和样本方差得 $9.02\leqslant\mu\leqslant10.98$。故在 95%的置信水平下该油层孔隙

度(总体)的期望值落在[9.02,10.98]这一范围内。

二、假设检验

1.正态分布总体期望值的假设检验

假设检验是除参数估计之外的另一类重要的统计推断问题,我们把任一关于总体分布的假设,统称为统计假设,简称假设。我们用关于正态分布总体期望值假设检验这一简单的情况,来说明假设检验的基本原理,在后面涉及的更复杂情况下的假设检验,其原理是相同的。

1)假设检验问题

关于正态分布总体期望值的假设检验问题是这样的:设有一个正态分布总体,其期望值为μ_0,且μ_0已知,方差为σ^2,方差已知或未知,我们先讨论最简单的情况,即假定方差σ^2已知,现在,我们有一个容量为n的样本,样本的均值为$\bar{Z}$,试问该样本是否能代表该总体或者说该样本是否来自于该总体?

回答上述问题,实际上是要回答样本的均值$\bar{Z}$与总体的期望值μ_0是否足够接近的问题,或者说样本所代表的真实总体的期望值μ与已知正态分布总体的期望值μ_0是否相等的问题,即确定是否定还是接受原假设(或称为零假设)

$$H_0: \mu = \mu_0$$

当原假设被否定时,则接受对立假设(或称为备择假设)

$$H_1: \mu \neq \mu_0 \text{ 或 } H_1: \mu > \mu_0 \text{ 或 } H_1: \mu < \mu_0$$

原假设和备择假设是相互对立的,两者有且只有一个正确。所谓假设检验问题,就是要判断原假设H_0是否正确,决定拒绝还是接受原假设,若拒绝原假设就接受备择假设。

根据备择假设H_1的不同表述,我们可以把假设检验分为双边假设检验、右单边假设检验和左单边假设检验三类,下面,我们以双边假设检验为例,讨论假设检验的规则。

2)假设检验规则

若原假设$H_0: \mu = \mu_0$为真,则$|\bar{Z} - \mu_0|$一般应较小;否则,$|\bar{Z} - \mu_0|$一般应较大。因此,我们可以根据$|\bar{Z} - \mu_0|$的大小,即差异是否显著来决定拒绝还是接受原假设。$|\bar{Z} - \mu_0|$愈大,我们愈倾向于拒绝原假设。那么$|\bar{Z} - \mu_0|$大到何种程度才能决定拒绝原假设呢?为此,需要制定一个检验规则:

当$|\bar{Z} - \mu_0| > C$时,拒绝原假设H_0;

当$|\bar{Z} - \mu_0| \leqslant C$时,接受原假设$H_0$。

其中,C是一个待定的常数,称为临界值。我们把拒绝原假设H_0的范围称为拒绝域,接受原假设H_0的范围称为接受域。因此,确定一个检验规则,事实上就是确定一个拒绝域(或接受域)。怎样确定拒绝域,这涉及假设检验中的两类错误问题。

由于样本具有随机性,根据样本做出判断就可能犯两类错误:

一类错误是原假设本来正确,但按检验规则却拒绝了原假设,这类错误称为弃真错误或第Ⅰ类错误,其发生的概率记为α,也就是说在原假设正确的情况下,由于样本的随机性引起的事件$\{|\bar{Z} - \mu_0| > C\}$出现的概率为$\alpha$,即

$$P\{|\bar{Z} - \mu_0| > C\} = \alpha \tag{2-32}$$

另一类错误是原假设本来不正确，但按检验规则却接受了原假设，这类错误称为取伪错误或第Ⅱ类错误，其发生的概率记为 β，也就是说在原假设本来不正确的情况下，由于样本的随机性引起的事件 $\{|\bar{Z}-\mu_0|\leqslant C\}$ 出现的概率为 β，即

$$P\{|\bar{Z}-\mu_0|\leqslant C\}=\beta \tag{2-33}$$

检验决策和两类错误的关系如表 2-6 所示。

表 2-6 检验决策和两类错误的关系

检验决策 \ H_0 状况	H_0 为真	H_0 非真
拒绝原假设 H_0	第Ⅰ类错误	正确
接受原假设 H_0	正确	第Ⅱ类错误

我们希望犯这两类错误的概率都非常小。但在一定的样本容量下，减小 α 会引起 β 增大，减小 β 会引起 α 增大。鉴于这种情况，Neyman 和 Pearson 提出一个原则，即在控制犯第Ⅰ类错误的概率 α 的条件下，尽量使犯第Ⅱ类错误的概率 β 小。这一原则的含义是，原假设要受到维护，使之不致轻易被否定，若检验结果否定原假设，说明否定的理由是充分的，同时，做出否定判断的可靠程度 $1-\alpha$ 也得到保证。所以，在实际应用中，为了通过样本观测值对某一陈述取得强有力的支持，通常把这种陈述本身作为原假设，而将对这一陈述的否定作为备择假设。在统计推断中，这种只控制 α 而不考虑 β 的假设检验，称为显著性检验，α 称为显著性水平，最常用的 α 值为 0.01，0.05，0.10 等。一般说来，根据研究问题的情况，若犯弃真错误损失大，为减小这类错误，α 取值小些，反之，α 取值大些。

在 α 足够小的情况下(一般也是这样选的)，犯弃真错误的概率就足够小，假设检验的基本思想可以用小概率原理来解释。所谓小概率原理，就是认为小概率事件在一次试验中是几乎不可能发生的，即在原假设 $H_0:\mu=\mu_0$ 为真的情况下，事件 $\{|\bar{Z}-\mu_0|>C\}$ 是几乎不可能发生的。也就是说，如果原假设 $H_0:\mu=\mu_0$ 是真的，那么不利于或不能支持这一假设的事件 $\{|\bar{Z}-\mu_0|>C\}$ 在一次试验中几乎不可能发生；倘若在一次试验中事件 $\{|\bar{Z}-\mu_0|>C\}$ 竟然发生了，我们有理由怀疑这一假设的真实性，并拒绝这一假设。

根据以上原理，在给定显著性水平(即弃真错误发生的概率)α 的情况下，若原假设 $H_0:\mu=\mu_0$ 为真，则有

$$P\{|\bar{Z}-\mu_0|>C\}=\alpha \tag{2-34}$$

同时，我们又知道

$$P\left\{\left|\frac{\bar{Z}-\mu_0}{\sigma/\sqrt{n}}\right|>z_{\alpha/2}\right\}=P\{|z'|>z_{\alpha/2}\}=\alpha \tag{2-35}$$

中的 $z_{\alpha/2}$ 可从标准正态分布表(附表 1)中读出。

显然，事件 $\{|\bar{Z}-\mu_0|>C\}$ 与事件 $\{|z'|>z_{\alpha/2}\}$ 是等同的事件，故也可把 $z_{\alpha/2}$ 称为临界值，此时的检验规则为：

当 $|z'|=\left|\frac{\bar{Z}-\mu_0}{\sigma/\sqrt{n}}\right|>z_{\alpha/2}$ 时，拒绝原假设；

当 $|z'| = \left|\dfrac{\bar{Z} - \mu_0}{\sigma/\sqrt{n}}\right| \leqslant z_{\alpha/2}$ 时，接受原假设。

由此，可确定拒绝域的面积为 α，接受域的面积为 $1-\alpha$（图 2 - 14）。

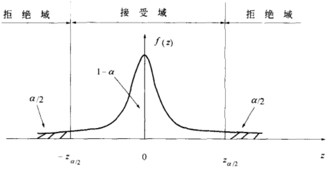


图 2 - 14　正态分布总体期望值标准正态分布双边检验中原假设接受域和拒绝域的分布

以上讨论的是双边检验的情况，下面对单边检验进行讨论。

当备择假设为 $H_1: \mu > \mu_0$ 时的检验称为右单边检验，检验规则如下：

当 $z' = \dfrac{\bar{Z} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$ 时，拒绝原假设；

当 $z' = \dfrac{\bar{Z} - \mu_0}{\sigma/\sqrt{n}} \leqslant z_\alpha$ 时，接受原假设。

图 2 - 15 是右单边检验中原假设接受域和拒绝域的分布示意图。

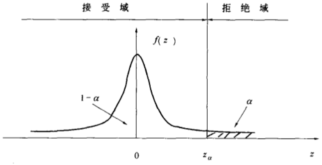


图 2 - 15　正态分布总体期望值标准正态分布右单边检验中原假设接受域和拒绝域的分布

当备择假设为 $H_1: \mu < \mu_0$ 时的检验称为左单边检验，检验规则如下：

当 $z' = \dfrac{\bar{Z} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$ 时，拒绝原假设；

当 $z' = \dfrac{\bar{Z} - \mu_0}{\sigma/\sqrt{n}} \geqslant -z_\alpha$ 时，接受原假设。

图 2 - 16 是左单边检验中原假设接受域和拒绝域的分布示意图。

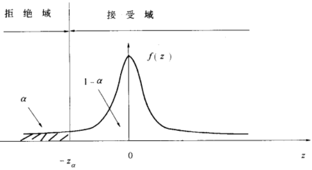


图 2-16　正态分布总体期望值标准正态分布左单边检验中
原假设接受域和拒绝域的分布

3)假设检验实例分析

【例 2-7】 某公司想从国外引进一种自动加工装置,该装置的工作温度 Z 服从正态分布 $N(\mu,5^2)$。厂方标明的平均工作温度是 80℃。从该装置试运行的 16 次随机测试得到的平均工作温度是 83℃。那么装置运行结果与厂方所标明的指标是否有差异呢？厂方的说法是否可以接受呢？

解:在该例子中,可以提出这样两个假设:一个称为零假设或原假设,记为 $H_0:\mu=80$℃;另一个称为备择假设或对立假设,记为 $H_1:\mu\neq80$℃。应如何做出判断呢？这里的问题是样本的平均工作温度 83℃,固然与厂方的指标 80℃有差异,但根据这种差异难以从直观上做出判断,无法决定接受还是拒绝厂方的说法。因为样本有随机性,80℃和 83℃的差异可能是样本的随机性造成的。在这种情况下,要对原假设 H_0 做出拒绝还是接受的判断,就必须根据研究的问题和决策条件,对样本值和原假设的差异进行分析,若有充分理由认为这种差异并非完全由随机因素造成的,即认为差异是显著的,才能拒绝原假设,否则不能拒绝原假设。

由于该装置的工作温度 Z 服从正态分布 $N(80,5^2)$,根据中心极限定理,样本容量为 $n=16$ 的样本的平均温度 $\bar{Z}$ 服从正态分布 $N(80,5^2/16)$,检验统计量为

$$|z'|=\left|\frac{\bar{Z}-80}{5/\sqrt{16}}\right|=2.4$$

在 $\alpha=0.05$ 的显著性水平下,读标准正态分布表(附表 1)得 $z_{\alpha/2}=1.96$。显然,$|z'|=2.4>z_{\alpha/2}=1.96$,故统计量$|z'|$值落在拒绝域(图 2-13),由此可以认为这种装置的实际平均工作温度与厂方说的有显著差异,故拒绝原假设 $H_0:\mu=80$。

4)正态分布总体期望值的 t 检验

t 检验法是用服从 t 分布的统计量检验正态总体期望值的方法。假设样本容量为 n,样本均值为 $\bar{Z}$,检验的目的是判断样本是否取自期望值为 μ_0(已知)和方差 σ^2(未知)的正态总体,即判断在 α 显著性水平下,样本是否源自该正态总体。

检验的本质实际上是看样本源自的期望值为 μ(未知)的真实总体与期望值为 μ_0(已知)的正态总体在 α 显著性水平下是不是同一个总体,即对假设 $H_0:\mu=\mu_0$ 的正确性进行检验,可分两种情况:(1)当 $n\geqslant30$ 时,在原假设 $H_0:\mu=\mu_0$ 为真的条件下,$\left|\frac{\bar{Z}-\mu_0}{s/\sqrt{n}}\right|$($s$ 为样本的均方差)近似服从标准正态分布,故可以用标准正态分布表进行假设检验,检验规则见表 2-7;

(2)当$n<30$时,在原假设 $H_0:\mu=\mu_0$ 为真的条件下,$\left|\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}\right|$ (s 为样本的均方差)服从自由度为 $n-1$ 的 t 分布(附表2),这时,用 t 分布表进行假设检验,检验规则见表2-7,图2-17为正态分布总体期望值 t 分布双边检验中原假设接受域和拒绝域的分布示意图。

表2-7 正态分布总体期望值的 t 检验表

检验类	H_0	H_1	$n\geqslant 30$	$n<30$
			在显著性水平 α 下拒绝 H_0,若	
I	$\mu=\mu_0$	$\mu\neq\mu_0$	$\left\|\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}\right\|>z_{\alpha/2}$	$\left\|\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}\right\|>t_{\alpha/2}(n-1)$
II	$\mu=\mu_0$	$\mu>\mu_0$	$\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}>z_\alpha$	$\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}>t_\alpha(n-1)$
III	$\mu=\mu_0$	$\mu<\mu_0$	$\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}<-z_\alpha$	$\dfrac{\bar{Z}-\mu_0}{s/\sqrt{n}}<-t_\alpha(n-1)$
备　注			$s=\sqrt{\sum_{i=1}^{n}(z_i-\bar{Z})^2/(n-1)}$	

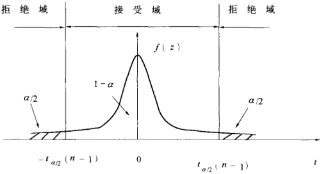


图2-17 正态分布总体期望值 t 分布双边检验中原假设接受域和拒绝域的分布

2.二正态分布总体期望值的假设检验

前面讨论了单正态总体的情况,对两个正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$,常常需要根据两个相互独立的样本对假设 $H_0:\mu_1-\mu_2=\delta$ (δ 是给定的一个实数,当然可以取0)进行检验,两个样本的平均值分别为 $\bar{Z}_1$ 和 $\bar{Z}_2$,方差分别为 s_1^2 和 s_2^2,容量分别为 n_1 和 n_2,下面分两种情况来讨论。

1)二总体的方差 σ_1^2 和 σ_2^2 已知

二总体方差已知时,在原假设 $H_0:\mu_1-\mu_2=\delta$ 为真的条件下;统计量 $(\bar{Z}_1-\bar{Z}_2-\delta)/\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$ 服从标准正态分布,于是,可以利用标准正态分布数据表(附表1)进行假设检验,检验准则见表2-8。

2)二总体方差未知但 $\sigma_1^2=\sigma_2^2$

二总体方差未知但 σ_1^2 与 σ_2^2 相等时,在原假设 $H_0:\mu_1-\mu_2=\delta$ 为真的条件下,统计量 $(\bar{Z}_1$

$-\bar{Z}_2-\delta)/(s_w\sqrt{1/n_1+1/n_2})$ 服从自由度为 (n_1+n_2-2) 的 t 分布，其中 s_w 为两个样本的平均均方差

$$s_w=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \tag{2-36}$$

式(2-36)中，$s_1=\sum_{i=1}^{n_1}(z_{i1}-\bar{Z}_1)^2/(n_1-1)$，$s_2=\sum_{i=1}^{n_2}(z_{i2}-\bar{Z}_2)^2/(n_2-1)$。

此时，可利用 t 分布数据表(附表2)进行假设检验，检验准则见表2-8。

表2-8　二正态分布总体期望值差的检验表

检验类	H_0	H_1	σ_1^2 和 σ_2^2 已知	σ_1^2 和 σ_2^2 未知，但 $\sigma_1^2=\sigma_2^2$
			在显著性水平 α 下拒绝 H_0，若	
1	$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2\neq\delta$	$\lvert\bar{Z}_1-\bar{Z}_2-\delta\rvert>\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}z_{\alpha/2}$	$\lvert\bar{Z}_1-\bar{Z}_2-\delta\rvert>s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha/2}(n_1+n_2-2)$
2	$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2>\delta$	$\bar{Z}_1-\bar{Z}_2-\delta>\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}z_\alpha$	$\bar{Z}_1-\bar{Z}_2-\delta>s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_\alpha(n_1+n_2-2)$
3	$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2<\delta$	$\bar{Z}_1-\bar{Z}_2-\delta<-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}z_\alpha$	$\bar{Z}_1-\bar{Z}_2-\delta<-s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_\alpha(n_1+n_2-2)$

3. 正态分布总体方差的检验——χ^2 检验

设样本 $z_1,z_2,\cdots,z_n$ 来自某一总体，检验该样本是否来自于方差为 σ_0^2(已知)的正态分布总体 $N(\mu,\sigma_0^2)$，即确定原假设 $H_0:\sigma^2=\sigma_0^2$ 是否为真，下面也分两种情况来讨论假设检验问题。

1)正态分布总体 $N(\mu,\sigma_0^2)$ 的期望值已知

当正态分布总体 $N(\mu,\sigma_0^2)$ 的期望值已知时，在原假设 $H_0:\sigma^2=\sigma_0^2$ 为真的条件下，统计量 ns^2/σ_0^2 服从自由度为 n 的 χ^2 分布，其中 $s^2=\sum_{i=1}^{n}(z_i-\mu)^2/n$ 为样本方差，可以用 χ^2 分布表(见附表3)进行检验，检验规则见表2-9。

表2-9　正态分布总体方差的 χ^2 检验表

检验类	H_0	H_1	μ 已知	μ 未知
			在显著性水平 α 下拒绝 H_0，若	
1	$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$ns^2/\sigma_0^2>\chi_{\alpha/2}^2(n)$ 或 $ns^2/\sigma_0^2<\chi_{1-\alpha/2}^2(n)$	$(n-1)s^2/\sigma_0^2>\chi_{\alpha/2}^2(n-1)$ 或 $(n-1)s^2/\sigma_0^2<\chi_{1-\alpha/2}^2(n-1)$
2	$\sigma^2=\sigma_0^2$	$\sigma^2>\sigma_0^2$	$ns^2/\sigma_0^2>\chi_\alpha^2(n)$	$(n-1)s^2/\sigma_0^2>\chi_\alpha^2(n-1)$
3	$\sigma^2=\sigma_0^2$	$\sigma^2<\sigma_0^2$	$ns^2/\sigma_0^2<\chi_{1-\alpha}^2(n)$	$(n-1)s^2/\sigma_0^2<\chi_{1-\alpha}^2(n-1)$
备　注			以上各行中 $s^2=\sum_{i=1}^{n}(z_i-\mu)^2/n$	以上各行中 $s^2=\sum_{i=1}^{n}(z_i-\bar{Z})^2/(n-1)$

2)正态分布总体 $N(\mu,\sigma_0^2)$ 的期望值未知

当正态分布总体 $N(\mu,\sigma_0^2)$ 的期望值未知时，在原假设 $H_0:\sigma^2=\sigma_0^2$ 为真的条件下，统计量

$(n-1)s^2/\sigma_0^2$ 服从自由度为$(n-1)$的 x^2 分布,其中 $s^2=\sum_{i=1}^{n}(z_i-\bar{Z})^2/(n-1)$为样本方差,可以用 χ^2 分布表(见附表3)进行检验,检验规则见表2-9。图2-18为正态分布总体方差 χ^2 双边检验原假设 $H_0:\sigma^2=\sigma_0^2$ 拒绝域和接受域的分布示意图,注意 χ^2 的变化范围为0~∝。

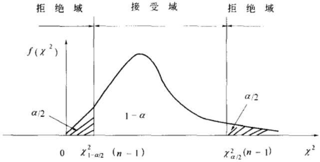


图2-18 正态分布总体期方差 χ^2 分布双边检验中原假设接受域和拒绝域的分布

4.二正态分布总体方差的检验——F 检验

设 $z_{11},z_{21},\cdots,z_{n_{1,1}}$ 为总体 $Z_1\sim N(\mu_1,\sigma_1^2)$的一个样本,$z_{12},z_{22},\cdots,z_{n_{2,2}}$ 为总体 $Z_2\sim N(\mu_2,\sigma_2^2)$的一个样本,并且两个样本相互独立,$F$ 检验即是根据样本的方差来判断总体 $Z_1\sim N(\mu_1,\sigma_1^2)$和总体 $Z_2\sim N(\mu_2,\sigma_2^2)$是否为同一总体,即判断原假设 $H_0:\sigma_1^2=\sigma_2^2$ 是否成立,同样可分两种情况来讨论。

1)当 μ_1 和 μ_2 已知时

当 μ_1 和 μ_2 已知时,统计量 $\sum_{i=1}^{n_1}(z_{i1}-\bar{Z}_1)^2/\sigma_1^2$ 服从 $\chi^2(n_1)$ 分布,统计量 $\sum_{i=1}^{n_2}(z_{i2}-\bar{Z}_2)^2/\sigma_2^2$ 服从 $\chi^2(n_2)$分布,在原假设 $H_0:\sigma_1^2=\sigma_2^2$ 为真的条件下,我们有

$$\frac{\sum_{i=1}^{n_1}(z_{i1}-\bar{Z}_1)^2/n_1\sigma_1^2}{\sum_{i=1}^{n_2}(z_{i2}-\bar{Z}_2)^2/n_2\sigma_2^2}=\frac{s_1^2}{s_2^2}\sim F(n_1,n_2) \tag{2-37}$$

因此,利用 F 分布数据表(附表4)即可进行二正态分布总体方差是否相等的假设检验,检验规则见表2-10。

表2-10 正态分布总体方差比的 F 检验表

检验类	H_0	H_1	μ_1 和 μ_2 已知	μ_1 和 μ_2 未知
			在显著性水平 α 下拒绝 H_0,若	
1	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$	$\frac{s_1^2}{s_2^2}>F_{\alpha/2}(n_1,n_2)$或 $\frac{s_1^2}{s_2^2}<F_{1-\alpha/2}(n_1,n_2)$	$\frac{s_1^2}{s_2^2}>F_{\alpha/2}(n_1-1,n_2-1)$或 $\frac{s_1^2}{s_2^2}<F_{1-\alpha/2}(n_1-1,n_2-1)$
2	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2>\sigma_2^2$	$\frac{s_1^2}{s_2^2}>F_{\alpha}(n_1,n_2)$	$\frac{s_1^2}{s_2^2}>F_{\alpha}(n_1-1,n_2-1)$
3	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2<\sigma_2^2$	$\frac{s_1^2}{s_2^2}<F_{1-\alpha}(n_1,n_2)$	$\frac{s_1^2}{s_2^2}<F_{1-\alpha}(n_1-1,n_2-1)$

续上表

检验类	H_0	H_1	μ_1 和 μ_2 已知	μ_1 和 μ_2 未知
			在显著性水平 α 下拒绝 H_0,若	
备　注			以上各行中 $s_1^2=\sum_{i=1}^{n_1}(z_{i1}-\mu_1)^2/n_1$ $s_2^2=\sum_{i=1}^{n_2}(z_{i2}-\mu_2)^2/n_2$	以上各行中 $s_1^2=\sum_{i=1}^{n_1}(z_{i1}-\bar{Z}_1)^2/(n_1-1)$ $s_2^2=\sum_{i=1}^{n_2}(z_{i2}-\bar{Z}_2)^2/(n_2-1)$

2)当 μ_1 和 μ_2 未知时

当 μ_1 和 μ_2 未知时,统计量 $\sum_{i=1}^{n_1}(z_{i1}-\bar{Z}_1)^2/\sigma_1^2$ 服从 $\chi^2(n_1-1)$分布,统计量$\sum_{i=1}^{n_2}(z_{i2}-\bar{Z}_2)^2/\sigma_2^2$ 服从 $\chi^2(n_2-1)$分布,在原假设 $H_0:\sigma_1^2=\sigma_2^2$ 为真的条件下,我们有

$$\frac{\sum_{i=1}^{n_1}(z_{i1}-\bar{Z}_1)^2/(n_1-1)\sigma_1^2}{\sum_{i=1}^{n_2}(z_{i2}-\bar{Z}_2)^2/(n_2-1)\sigma_2^2}=\frac{s_1^2}{s_2^2}\sim F(n_1-1,n_2-1) \tag{2-38}$$

因此,利用 F 分布数据表(附表 4)即可进行二正态分布总体方差是否相等的假设检验,检验规则见表 2-10。图 2-19 为正态分布总体方差比的双边 F 检验原假设 $H_0:\sigma^2=\sigma_0^2$ 拒绝域和接受域的分布示意图,注意 F 的变化范围为 $0\sim\infty$。

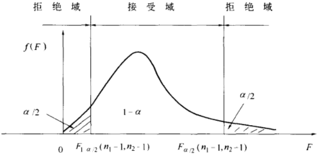


图 2-19　正态分布总体期方差比的 F 分布双边检验中原假设接受域和拒绝域的分布

在二正态分布总体方差是否相等的 F 检验中,F 分布的上 100α 百分点 $F_\alpha(n_1,n_2)$是指满足 $\int_{F_{\alpha(n_1,n_2)}}^{+\infty}f(F)\mathrm{d}F=\alpha(0<\alpha<1)$ 的点,其中 $f(F)$是 F 分布的概率密度,$F_\alpha(n_1,n_2)$的值可由 F 分布表(附表 4)查到。对于表中不能查到的值,可用关系式 $F_{1-\alpha}(n_1,n_2)=\frac{1}{F_\alpha(n_2,n_1)}$得到。

第五节　概率分布函数的拟合与应用

一、概率分布函数拟合过程

在具体研究一个实际变量时,如油气藏储量规模,我们常常是有一系列数据(样本),该变

量服从什么分布，我们并不知道，这就需要从实际数据出发，进行分布函数拟合，以找到该变量或者说由数据揭示的变量分布的概率模型。

概率分布函数拟合可以借助计算机程序进行，而简便和实用的办法是借助于概率纸来完成操作，下面以表 2－11 中某区气藏规模分布数据概率分布模型的拟合来说明操作步骤：

表 2－11　某区气藏规模分布数据及“大于”概率计算

编　号	规模，bcf	排序(从大到小)	“大于”概率，%
1	150	320	8.33
2	20	150	16.67
3	40	90	25
4	5	70	33.33
5	15	55	41.67
6	320	40	50
7	10	30	58.33
8	30	20	66.67
9	90	15	75
10	55	10	83.33
11	70	5	91.67

(1)把变量的样本值列于表中(表 2－11 中的第 2 列)；

(2)把样本值按从大到小的顺序排序(表 2－11 中的第 3 列)(在油气地质应用中，一般我们更关心的是某变量大于某个值的概率，故把样本值按从大到小的顺序排序，当我们更关心某变量小于某个值的概率时，把样本值按小到大的顺序排序)；

(3)计算“大于”概率值 $AF(z) = P\{Z > z\} = \frac{i}{n+1}$(见表 2－11 中的第 4 列)，其中，$i$ 为序号，n 为样本容量(当第(2)步中的样本值按从小到大的顺序排序时，计算得到的是“小于”概率值 $F(z) = P\{Z \leqslant z\} = \frac{i}{n+1}$)；

(4)把变量值 z 和其对应的“大于”概率值 $P\{Z > z\}$ 投到对数正态分布概率纸上(可以是任何分布模型的概率纸，在这里之所以选择对数正态分布概率纸，是因为大量经验表明，油气藏储量规模常常服从对数正态分布)，若得到一条直线(图 2－16)，则变量服从对数正态分布，否则，尝试使用其它分布类型的概率纸。

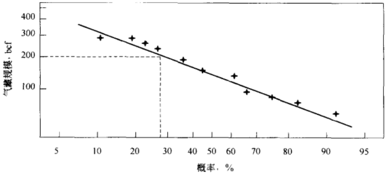


图 2－20　变量分布函数的拟合(示意图)

注意图 2－20 仅为示意图，不对应表 2－11 中的实际数据。

二、利用某种分布概率纸进行拟合的基本原理

到此，读者可能会问，为什么用概率纸就可以达到概率分布模型拟合的目的？我们现在来讨论这个问题。

首先，我们来看一看直接把 $z_i—AF(z_i)$ 投到算术坐标系中会是什么情形，图 2－21 中过各点连起来的曲线实质上是气藏规模的累积概率分布曲线（经验），根据这条曲线来判别气藏规模服从什么分布是很困难的，因为曲线缺少直观性。如果我们将坐标系适当刻度，使图 2－21 中的曲线变成直线，那么，问题就变得简单了。

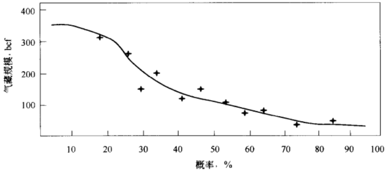


图 2－21　气藏规模的累积概率分布函数（示意图）

假定变量 Z（如气藏规模）服从对数正态分布，意味着变量 $\ln Z$ 服从正态分布，则有

$$\frac{\ln Z - \mu}{\sigma} = Z' \sim N(0,1) \tag{2-39}$$

式（2－39）表明，$\ln Z$ 与标准正态分布变量 Z' 之间存在线性关系，而每个 Z' 值对应惟一一个 $AF(z') = P\{Z' > z'\}$。若纵坐标 Z 取对数坐标系，横坐标 F 用对应 z' 的 $AF(z')$ 来刻度，则在经过刻度的坐标系中点出 $z_i—AF(z_i)$，把这些点连起来，必然会得到一条直线（图 2－20）。由于直线非常直观，易于用来判别数据是否服从假定的分布，所以，利用分布概率纸进行拟合成为常用的方法。

三、概率分布函数的应用

通过前面的步骤，得到的直线实际上是变量概率分布函数在适当刻度坐标系中的概率分布函数，利用该直线，就可方便地读出给定的概率值所对应的变量值或给定的变量值所对应的概率值。从分布直线上可直接读出变量值大于某一给定值发生的概率，如对应 ＞200bcf 的气藏规模，从图 2－20 可示意性地读出其对应的概率为 28％ 或 0.28，也就是说，发现储量规模 ＞200bcf 气藏的概率或可能性为 0.28。

◆ 思 考 和 练 习 题 ◆

1．如何理解随机变量和随机向量？

2．有两个随机变量 Z_1 和 Z_2（见下表），求其联合概率密度，并判别这两个变量是否相互独立。

两个随机变量取值表

Z_1	Z_2	Z_1	Z_2	Z_1	Z_2
3	6	1	10	2	9
2	7	2	9	4	8
2	8	2	10	1	12
2	8	1	11	5	6

3.如何理解中心极限定理？哪些条件下中心极限定理适用？哪些条件下中心极限定理不适用？

4.取 n 块岩心测得孔隙度 ϕ 和渗透率 K，求得 $\log K$ 与 ϕ 之间的线性相关系数为 0.62，问当 n 分别为 9、16、36、64 时，在 5%的显著性水平下，$\log K$ 与 ϕ 之间是否显著相关？从 n 的变化中能得出什么结论？

5.已知某油层孔隙度服从正态分布，在该油层中取 25 块岩心作为样本，测得其孔隙度，得到样本均值 $\bar{Z}=10$，样本方差 $s^2=\frac{\sum_{i=1}^{n}(z_i-\bar{Z})^2}{n-1}=9$，问在 95%的置信水平下，该油层孔隙度（总体）的期望值落在哪个区间内？

6.已知某油层孔隙度 Z 服从正态分布 $N(15,3^2)$。从该油层中取 36 块岩心，36 块岩心分析数据表明孔隙度均值为 20%，问这 36 块岩心能否代表该油层？

7.用表 2-11 中的数据，在对数正态分布概率纸上绘出概率分布线，并求出油气藏规模大于 50bcf 的概率有多大？假设该区的勘探成功率为 25%，问在该区找到一个大于 50bcf 气藏的概率有多大？

第三章　相关、回归及其应用

本章学习目的和要求：了解相关、一元线性回归和多元线性回归的基本概念，理解线性相关和一元线性回归的联系和区别以及各自的分析方法，熟练掌握最小二乘法方法原理、应用条件及其在一元线性回归和多元线性回归分析中的应用。

研究客观事物的相互关系，既要做定性分析，也要做定量分析，考察它们的联系程度，揭示其变化的具体形式和变化规律，相关和回归分析便是这种定量分析的重要统计方法。在自然科学、工程技术和经济领域都有着广泛的应用。地学研究领域中，在研究某一地质因素与特定地质现象的关系时，相关和回归分析已成为关系模型建立、结构分析和预测分析的重要工具。

本章首先通过介绍简单线性相关和一元线性回归，阐明相关和回归的基本原理及它们的相互联系和区别，其次展开分析如何利用最小二乘方法进行一元线性回归、多元线性回归分析及回归分析结果的假设检验。

第一节　线 性 相 关

一、相关关系和相关系数

1.相关

自然界中许多现象之间存在着相互依赖和相互制约关系。这种关系表现在量上主要有两种类型：一类是函数关系；另一类是统计关系或称相关关系。相关关系是指变量之间的不确定的依存关系。它同通常的确定的函数关系不同。通常的函数关系表示变量之间确定的依存关系，对给定的自变量，有一个相应的因变量数值，如圆的面积 $S=\pi r^2$，给定半径 r 便有一个确定的面积 S。但是由于地质问题的复杂性，大部分地质变量都是随机变量，它们之间的关系大多属于相关关系，不能用确定的数学函数表示，而只能用某种统计关系表示。例如，油气储层中的孔隙度与渗透率之间的关系和有机质总量与泥岩体积之间的关系等。在相关关系中，变量之间既具有某种依赖关系，又具有某种不确定性，即不能根据自变量的值，将因变量的值完全确定下来。

由于客观事物的联系和变化复杂多样，变量间的相关关系也有多种形式。

按研究变量的多少可划分为一元相关（也称为简单相关）和多元相关（也称为复相关）：两个变量的相关关系称为一元相关。如总的生烃量与泥岩中有机质含量之间的关系和声波时差与孔隙度的关系等。三个或三个以上变量的相关关系称为多元相关，如某圈闭的油气储量与孔隙度、圈闭面积、圈闭有效幅度和含油气饱和度的关系等。

按研究变量之间的依存关系的形式划分为线性相关（也称直线相关）和非线性相关（也称曲线相关）：当一个变量每减少一个单位，另一个相关变量按一个大致固定的增（减）量变化时称为线性相关；反之，相关变量不按固定的增（减）量变化时称为非线性相关。例如，大多数地层的泥质含量与自然放射性强度有线性相关关系，而油气储集层的孔隙度和渗透率的关系通常为非线性关系。

按变量变化的方向划分为正相关和负相关:相关的变量按同一方向变化,即一个变量由小到大或由大到小的变化时,相关变量随之由小到大或由大到小变化,为正相关;相关的变量按相反方向变化,即一个变量由小到大或由大到小的变化时,另一变量却由大到小或由小到大变化,为负相关。例如,河道砂的储层厚度与砂体的延伸长度有正相关关系,而泥质含量与储层孔隙度的关系是负相关关系。

此外,相关关系还可以按变量之间关系的密切程度划分为完全相关、不相关(或零相关)和不完全相关:当变量之间的依存关系密切到近于函数关系时,称为完全相关;当变量之间不存在依存关系时,就称为不相关或零相关;大多数相关关系介于其间,称为不完全相关。

2.相关系数

简单线性相关是两个变量的变化表现为直线形式的相关关系。简单相关分析就是对来自同一总体的两个变量之间的相关关系进行分析,其目的是对变量之间是否存在必然的联系、联系的形式、变化的趋势做出符合实际的判断,并考察它们联系的密切程度,检验其有效性。

早在 1890 年,英国统计学家 Karl Pearson 便提出了一个测定两个变量之间线性相关的计算公式,通常称为积矩相关系数或动差相关系数,其计算公式为

$$r = \frac{\sigma_{xz}}{\sigma_x \sigma_z} \tag{3-1}$$

式中,σ_{xz}是 x 与 z 变量的协方差;σ_x 是 x 变量的标准差;σ_z 是 z 变量的标准差。

把协方差的计算公式(见第二章)代入式(3-1)得到

$$r = \frac{\sum (x_i - \bar{x})(z_i - \bar{z})}{n\sigma_x \sigma_z} \tag{3-2}$$

式中,$\bar{x}$ 和 $\bar{z}$ 分别是 x 变量 z 变量观测值的平均值;n 为样品个数。

式(3-2)适用于实测样本数据,将式(3-2)稍做变换

$$r = \frac{\sum \frac{(x - \bar{x})}{\sigma_x} \frac{(z - \bar{z})}{\sigma_z}}{n} \tag{3-3}$$

即可明显看出相关系数的一些性质:

(1)相关系数是一个系数,不受变量值大小和计量单位的影响,便于在不同资料之间对相关程度进行比较;

(2)相关系数 r 的数值有一定的范围,即$|r| \leqslant 1$。当$|r| = 1$ 时,表示 x 与 z 变量为完全的线性相关,也即为确定的函数关系;当$|r| = 0$ 时,表示 x 与 z 变量不存在线性相关;$0 < |r| < 1$ 时,表示两个变量存在不同程度的线性相关。由此,可以确定一个对相关程度评价的标准。

3.相关系数的计算

1)未分组资料相关系数的计算

对于给定的 n 组测量数据(x_i, z_i),$i = 1,2\cdots n$,则上述的相关系数的公式可以表示为

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (z_i - \bar{z})^2}} \tag{3-4}$$

2)分组资料相关系数的计算

当原始数据较多,已分组编成二元频数分布表(即联合概率分布)时,就要用各组的频数加

权计算相关系数，计算公式为

$$r=\frac{n\sum_{i=1}^{n}x_iz_if_{x_iz_i}-\left(\sum_{i=1}^{n}x_if_{x_i}\right)\left(\sum_{i=1}^{n}z_if_{z_i}\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2f_{x_i}-\left(\sum_{i=1}^{n}x_if_{x_i}\right)^2}\times\sqrt{n\sum_{i=1}^{n}z_i^2f_{z_i}-\left(\sum_{i=1}^{n}z_if_{z_i}\right)^2}} \tag{3-5}$$

式中，f_{x_i} 是 x_i 的频数；f_{z_i} 是 z_i 的频数；$f_{x_iz_i}$ 是 x_i 与 z_i 的联合频数。

二、相关系数的抽样分布和显著性检验

相关系数既可以根据总体的全部资料计算，也可以根据随机抽样取得的部分资料计算，然后推断总体的相关系数。然而，在实际工作中，从一个二元正态总体中抽取一个样本集，由于受抽样误差的影响，在做出两个总体变量相关的结论前，必须检验样本相关系数 r 的显著性。事实上，要检验总体中的两个变量是否相关，即要检验两个变量间的相关系数 $r=0$ 或 $r\neq0$，前者表示两个变量不相关，后者表示两个变量具有相关性。

在小样本的情况下，上述相关性检验可用 Fisher 的 t 检验实现，方法原理参见第二章的有关内容，下面仅介绍检验的步骤。

1)提出假设

$H_0: r=0; H_1: r\neq0$，以 $|t|=|r|\sqrt{\dfrac{n-2}{1-r^2}}$ 为检验统计量，其中 n 为样品个数。

2)查 $t_{\alpha/2}(n-2)$ 的临界值

根据给定的显著性水平 α，查 t 分布表(附表 2)中 $t_{\alpha/2}(n-2)$ 的临界值。

3)比较计算的 $|t|$ 与 $t_{\alpha/2}(n-2)$ 临界值

比较计算的 $|t|$ 与 $t_{\alpha/2}(n-2)$ 临界值，如 $|t|$ 大于临界值，就拒绝原假设；如 $|t|$ 小于临界值，就接受原假设。

这里需要说明的是，上述相关系数仅限于对简单线性相关关系的测定，$r=0$ 也只表示两个随机变量不存在线性相关，而不能说明两变量为独立变量，因为在这种情况下，两个变量仍然可能存在非线性相关关系。另外，简单线性相关除了用上述相关系数表示外，也可通过一元线性回归模型的拟合优度系数表示。

第二节　简单线性回归

一、回归分析的基本概念

“回归”(Regression)一词源于 19 世纪英国生物学家 Galton 对人体遗传特征的实验研究。他根据实验数据发现个子高的双亲其子女也较高，但平均地看，个子特高的双亲其子女的身高不比他们的双亲高；同样，发现个子矮的双亲其子女也较矮，但平均地看，个子特矮的双亲其子女的身高不比他们的双亲矮。Galton 把这种身高趋向于人的平均高度的现象称为“回归”(Regression)，并将该概念加以应用。后来，他又提出了“相关”(Correlation)的概念，由此逐步形成具有独特理论和方法体系的回归分析。但是，现在统计学中的“回归”概念已不是原来生物学上的特殊规律性，而是指变量间的依存关系。

回归分析和相关分析都是对客观事物数量依存关系的分析，在理论基础和方法上具有一

致性。只有存在相关关系的变量才能进行回归分析，相关程度愈高，回归分析的结果愈可靠。因此，相关系数也是判定回归效果的一个重要依据。另一方面，相关系数同回归模型中的参数可以相互换算，特别是多元相关和非线性相关的相关系数必须利用回归模型才能求得。但是，回归分析和相关分析也有如下差别：

第一，相关分析是研究变量之间的共变关系，这些变量相互对应，却不分主从或因果。回归分析却是在控制或给定一个或几个变量条件下来观察某一对应变量的变化，给定的变量称为自变量。自变量可以是随机变量，也可以是确定型变量，一般假定为确定型变量。被观察的对应变量称为因变量，是随机变量。当给定一个自变量时，因变量可能有几个取值，并且形成一个分布。因此，回归分析必须根据研究目的和对象的性质，确定哪个是自变量，哪个是因变量。

第二，相关分析主要测定变量之间关系的密切程度和变量变化的方向，而回归分析可以对具有相关关系的变量建立一个数学方程，也称回归模型，用于描述变量之间具体的变动关系，通过控制或给定自变量的数值来估计或预测因变量可能的数值。

回归分析根据实际资料建立的回归模型也有多种形式。按自变量的多少可分为一元回归模型和多元回归模型；按变量之间的具体变动形式可以分为线性回归模型和非线性回归模型。把这两种分类标志结合起来，就有一元线性回归模型和一元非线性回归模型、多元线性回归模型和多元非线性回归模型。

二、简单线性回归模型

1.一元线性回归的数学模型

具有相关关系的变量之间虽然不具有确定的函数关系，但可借助回归函数关系表达它们的统计规律。回归函数是用于近似描述具有相关关系的变量间联系的函数。

设有一个自变量 x 和一个因变量 z，对于任意给定的 x_i，简单线性回归模型为

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \forall i \tag{3-6a}$$

式中，x_i——自变量，它可以是随机变量，也可以是数值可控制的确定型变量；

z_i——因变量，是一个随机变量；

β_0, β_1——统称为回归模型的参数，也称为回归系数；

ε_i——随机变量，且 $\varepsilon_i \sim N(0,\sigma^2)$；

$\forall i$——对于任意 i。

方程(3 - 6a)在平面直角坐标系中表示一条直线，回归分析中称之为回归直线。回归分析的主要目的是估计式(3 - 6a)中的参数 β_0, β_1，进而可通过给定的 x 来估计 z。那么，这里就涉及如下一系列问题：回归模型是否合适？估计精度如何？怎样进行判断和检验？下面从理论和计算两个方面讨论这些问题。

2.一元线性回归系数估计——最小二乘法基本原理

根据 n 对数据$(x_i, z_i)(i = 1,2,\cdots,n)$求 x、z 之间的回归方程，即需要求出回归模型(3 - 6a)的回归系数 β_0, β_1。

设 $z_i(i = 1,2,\cdots,n)$的估计值表示为

$$\hat{z}_i = b_0 + b_1 x_i (i = 1,2,\cdots,n) \tag{3-6b}$$

式中，b_0, b_1 是式(3 - 6a)中回归系数 β_0, β_1 的估计值。由于实际测定的对应于 x_i 的某个数值的

z_i 可能有多个值，通过 x_i 与 z_i 的各数值也可能有多条直线。无疑其中最具代表性的应该是实际值同这条直线平均离差最小的直线，亦即满足 $\sum_{i=1}^{n}(z_i - \hat{z}_i)^2 = \min$ 的直线。在满足这一要求的条件下，求出回归系数 β_0, β_1 的估计值 b_0, b_1，这一方法称为最小二乘法，其数学表达式为

$$Q = \sum_{i=1}^{n}(z_i - \hat{z}_i)^2 = \sum_{i=1}^{n}[z_i - (b_0 + b_1 x_i)]^2 = \min \tag{3-7}$$

根据微分学中求极值的原理，式(3-7)可写成如下形式

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n}[z_i - (b_0 + b_1 x_i)] = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n}[z_i - (b_0 + b_1 x_i)]x_i = 0 \end{cases} \tag{3-8}$$

式(3-8)经整理后得到

$$\begin{cases} nb_0 + \sum_{i=1}^{n} x_i b_1 = \sum_{i=1}^{n} z_i \\ \sum_{i=1}^{n} x_i b_0 + \sum_{i=1}^{n} x_i^2 b_1 = \sum_{i=1}^{n} x_i z_i \end{cases} \tag{3-9a}$$

解方程(3-9a)得到

$$\begin{cases} b_1 = \dfrac{\sum_{i=1}^{n} x_i z_i - n\bar{x}\bar{z}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ b_0 = \bar{z} - b_1 \bar{x} \end{cases} \tag{3-9b}$$

这样就得到了回归系数 β_0, β_1 的估计值 b_0, b_1，从而利用式(3-6b)，就可得到对应于给定自变量值 x_0 的 z 变量的估计值 $\hat{z}_0 = b_0 + b_1 x_0$。

三、拟合优度系数 r^2 和估计标准误差

1. 拟合优度系数 r^2

前面用最小二乘方法求得的回归直线 $\hat{z}_i = b_0 + b_1 x_i$ 确定了 x 和 z 的具体变化关系。但是，实际值是否紧密分布在该直线的两侧呢？其紧密程度如何？这些问题关系到回归模型的应用价值，因此，对回归直线的拟合优度必须加以测定。下面讨论拟合优度系数 r^2 的确定方法。

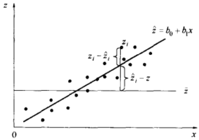


图 3-1　离差 $(z_i - \bar{z})$ 分解示意图

设有一个 z 的实际值 z_i，z_i 到 $\bar{z}$ 的离差即 $(z_i - \bar{z})$ 被回归直线分解成 $(z_i - \hat{z}_i)$ 和 $(\hat{z}_i - \bar{z})$ 两部分(图 3-1)，即有

$$(z_i - \bar{z}) = (z_i - \hat{z}_i) + (\hat{z}_i - \bar{z}) \tag{3-10}$$

对所有的实际值 z_i ($i = 1, 2, \cdots, n$)，求离差平方和，可得如下关系式

$$\sum_{i=1}^{n}(z_i - \bar{z})^2 = \sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2 + \sum_{i=1}^{n}(z_i - \hat{z}_i)^2 \tag{3-11}$$

式中，$\sum_{i=1}^{n}(z_i - \bar{z})^2$ 为总偏差；$\sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2$ 为回归偏差，且 $\sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2 = b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2$，由此可见回归偏差很大程度上取决于回归系数，所以也被称为回归解释的偏差；$\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$ 为剩余偏差，也称为未被解释的偏差。

式(3-11)表明，z 的实际值同 z 的均值的总偏差包括两部分：一部分是回归偏差，即 z 和 x 依存关系影响的偏差；另一部分，是各种不确定因素引起的随机误差。在总偏差一定时，回归偏差愈大，剩余偏差就愈小；反之，回归偏差愈小，剩余偏差就愈大。由此，如果 z 的实际值 $z_i(i=1,2,\cdots n)$ 都紧密分布在回归直线的两侧，剩余偏差就很小，说明 z 和 x 的依存关系很强，总偏差 $\sum_{i=1}^{n}(z_i - \bar{z})^2$ 主要由回归偏差 $\sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2$ 解释。在极端情况下，如 z_i 都落在回归直线上，$\sum_{i=1}^{n}(z_i - \hat{z}_i)^2 = 0$，则 $\sum_{i=1}^{n}(z_i - \bar{z})^2 = \sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2$，说明 z 和 x 为确定的函数关系。这时总偏差 $\sum_{i=1}^{n}(z_i - \bar{z})^2$ 就完全由回归偏差解释了。拟合优度系数 r^2 以回归偏差占总偏差的比率来表示回归模型拟合的优劣程度，其计算公式为

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3-12}$$

当 z 和 x 两个变量的线性依存关系很密切乃至 z 的变化完全由 x 引起时，$r^2=1$；当 z 和 x 两个变量不存在线性相关关系时，$r^2=0$；一般情况下，r^2 处于 0 和 1 之间。在实际应用中，由实测数据计算得到的回归系数可直接用于计算拟合优度系数

$$r^2 = \frac{b_0\sum_{i=1}^{n}z_i + b_1\sum_{i=1}^{n}x_iz_i - n\bar{z}^2}{\sum_{i=1}^{n}z_i^2 - n\bar{z}^2} \tag{3-13}$$

从上面的分析和拟合优度系数的定义可以看到，拟合优度系数 r^2 和前面讨论的相关系数具有一致性。事实上，可以证明(留给读者自行完成)一元线性回归的拟合优度系数 r^2 的平方根，就是简单线性相关的相关系数 r。

$$r = \pm\sqrt{\frac{\sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}} \tag{3-14}$$

式中，正负号应根据回归模型中的回归系数 β_1 的符号来确定。

2. 估计标准误差

如上所述，z 的实际值同 $\bar{z}$ 的总偏差中，回归偏差和剩余偏差是此长彼消的关系。与回归偏差可以从正面来衡量线性模型的拟合优度一样，剩余偏差可以从反面来判定线性模型的拟合优度。剩余偏差 $\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$ 除以自由度($n-2$)所得之商的平方根为估计标准误差，其公式为

$$s_{xz} = \sqrt{\frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{n-2}} \tag{3-15}$$

在回归分析中,估计标准误差愈小,表明估计值愈紧靠实际值,回归模型拟合得愈好;反之,估计标准误差愈大,则说明实际值对估计值愈分散,回归模型拟合得愈差。实际应用中估计标准误差可采用如下的简化公式

$$s_{xz} = \sqrt{\frac{\sum_{i=1}^{n} z_i^2 - b_0 \sum_{i=1}^{n} z_i - b_1 \sum_{i=1}^{n} x_i z_i}{n-2}} \tag{3-16}$$

作为回归模型拟合优度的判定指标,估计标准误差显然不如拟合优度系数,因为拟合优度系数是一个无量纲的系数,有确定的取值范围,便于对不同资料的回归模型的拟合优度进行比较,而估计标准误差是有计量单位的,又没有确定的取值范围,不便于对不同资料回归模型拟合优度进行比较。尽管如此,估计标准误差在回归分析中仍是一个重要的指标,它是用自变量估计因变量时,确定置信区间的尺度。

在大样本的条件下($n \geqslant 30$),用 x_0 对 z_0 进行估计的置信区间为

$$z_0 = \hat{z}_0 \pm z_{\alpha/2} s_{xz} = (b_0 + b_1 x_0) \pm z_{\alpha/2} s_{xz} \tag{3-17}$$

在小样本的情况下($n < 30$),要用 t 分布来确定置信区间。在给定的置信水平 $1-\alpha$ 时,z_0 的置信区间为

$$z_0 = \hat{z}_0 \pm t_{\alpha/2}(n-2) s_{xz} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{3-18}$$

式中,$t_{\alpha/2}(n-2)$ 可查 t 分布表(附表 2)得到,x_0 为给定的自变量的某一数值。

四、一元线性回归模型的显著性检验

从总体中随机抽取一个样本,根据样本的 n 对 x 和 z 的数据求得的线性回归模型,由于受抽样误差的影响,它确定的变量之间的线性关系是否显著及用该回归模型对给定的自变量 x 估计的因变量 z 是否有效,必须通过显著性检验才可得出结论。一元线性回归模型的显著性检验包括回归系数 β_1 的检验和模型整体的 F 检验。

1. 回归系数 β_1 的检验

回归系数 β_1 是决定 z 和 x 变量线性依存关系的重要参数。如果 $\beta_1 = 0$,表示 z 和 x 之间不存在线性关系,因此,检验总体回归系数 $\beta_1 = 0$ 的假设等于检验总体 z 和 x 变量没有线性关系的假设。

$\beta_1 = 0$ 的检验统计量为

$$S = \frac{b_1 - \beta_1}{\sigma_b} = \frac{b_1}{\sigma_b} \tag{3-19}$$

式中,统计量 S 服从均值为 0、方差为 1 的标准正态分布,σ_b 是样本回归系数 β_1 的标准误差,通常是未知的,要用它的估计量 $\hat{\sigma}_b$

$$\hat{\sigma}_b = \sqrt{\frac{s_{xz}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\left(\sum_{i=1}^{n} z_i^2 - b_0 \sum_{i=1}^{n} z_i - b_1 \sum_{i=1}^{n} x_i z_i\right)/(n-2)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \tag{3-20}$$

代替，其中 s_{xz} 为标准估计误差，见式(3-15)。

在大样本条件下，按此统计量 S 对 β_1 进行显著性检验，可查标准正态概率表(附表1)确定临界值。如果是小样本，则须用标准化处理的 t 变量，即

$$|t| = \frac{|b_1 - \beta_1|}{\hat{\sigma}_b} = \frac{|b_1|}{\hat{\sigma}_b} \tag{3-21}$$

式(3-21)中的 t 变量服从自由度为 $(n-2)$ 的 t 分布，可以查 t 分布表(附表2)确定临界值 $t_{\alpha/2}(n-2)$。通过对比实际计算的 $|t|$ 值和查表得到的临界值 $t_{\alpha/2}(n-2)$，若 $|t| \leqslant t_{\alpha/2}(n-2)$，则接受 $\beta_1=0$ 这一假设；若 $|t| > t_{\alpha/2}(n-2)$ 则拒绝 $\beta_1=0$ 这一假设(即接受备择假设 $\beta_1 \neq 0$)。

2.模型整体的 F 检验

回归模型的 F 检验是将总偏差 $\sum_{i=1}^{n}(z_i-\bar{z})^2$ 进行分解的一种检验方法。式(3-11)中的各种偏差都与一个自由度相联系：总偏差的自由度为 $(n-1)$，因为在计算 $\sum_{i=1}^{n}(z_i-\bar{z})^2$ 时消失了一个自由度；回归偏差的自由度为1，对一元线性回归模型来说，只有一个自变量与因变量对应；剩余偏差的自由度为 $(n-2)$，而且这三种偏差的自由度存在如下关系

$$(n-1) = 1 + (n-2) \tag{3-22}$$

将回归偏差和剩余偏差各自除以它们的自由度后再相除，便得到检验统计量 F，即

$$F = \frac{\sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2/1}{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2/(n-2)} \tag{3-23}$$

对比统计量 F 与拟合优度系数 r^2，可以用拟合优度系数表示统计量 F，即

$$F = \frac{r^2(n-2)}{1-r^2} \tag{3-24}$$

对总体回归系数 $\beta=0$ 的假设检验，是在给定显著性水平 α 条件下，将计算的统计量 F 与查表(附表4)得到的临界值 $F_{\alpha/2}(1,n-2)$ 或 $F_{1-\alpha/2}(1,n-2)$ 比较，若 $F < F_{\alpha/2}(1,n-2)$ 或 $F > F_{1-\alpha/2}(1,n-2)$，就接受原假设 $\beta=0$；若 $F > F_{\alpha/2}(1,n-2)$ 或 $F < F_{1-\alpha/2}(1,n-2)$，就拒绝原假设(即接受备择假设 $\beta \neq 0$)。

五、一元线性回归分析实例

【例3-1】 以家庭为单位，某种商品的需求量与该商品的价格之间的一组调查数据如表3-1所示，求价格对需求量的回归关系。

表3-1 商品价格与需求量实际调查数据

价格 x,元	5.0	2.0	2.0	2.3	2.5	2.6	2.8	3.0	3.3	3.5
需求量 z,斤	1.0	3.5	3.0	2.7	2.4	2.5	2.0	1.5	1.2	1.2

解：第一步：把测量数据点到平面直角坐标系上，得到散点图，考察相关关系；

第二步：根据点的分布趋势，选择式(3-6b)作为回归模型；

第三步：计算回归系数 b_0, b_1 根据表3-1的数据，计算出式(3-9a)中所需的各项数值并列于表3-2中；

表 3-2 回归模型参数计算中中间变量的数值

x_i	5.0	2.0	2.0	2.3	2.5	2.6	2.8	3	3.3	3.5
z_i	1.0	3.5	3.0	2.7	2.4	2.5	2.0	1.5	1.2	1.2
x_iz_i	5.0	7.0	6.0	6.0	6.5	5.6	4.5	4.0	4.2	55.0
x_i^2	1.0	4.0	4.0	6.3	6.8	7.8	9.0	10.9	12.3	67.3

第四步：得到回归系数 $b_1 = \dfrac{\sum_{i=1}^{n} x_i z_i - n\bar{x}\bar{z}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = -1.6, b_0 = \bar{z} - b_1\bar{x} = 6.5$，所求回归方程为 $z = 6.5 - 1.6x$。

第五步：检验回归方程的显著性，读者自行完成。

第三节　多元线性回归

一、多元线性回归的数学模型和回归方程

设有 m 个自变量 $x_1, x_2, \cdots\cdots, x_m$ 和一个因变量 z，用自变量来表示因变量的回归模型可写成如下形式

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots\cdots + \beta_m x_{im} + \varepsilon_i \quad \forall i \tag{3-25}$$

式中，$\beta_0, \beta_1, \beta_2, \cdots, \beta_m$ 是待定常数；z_i 和 ε_i 都是随机变量，各自服从不同期望值的正态分布 $z_i \sim N(\mu, \sigma^2)$ 和 $\varepsilon_i \sim N(0, \sigma^2)$，且 ε_i 之间是相互独立的；自变量 $x_1, x_2, \cdots\cdots, x_m$ 可以是随机变量，也可以是数值可控制的确定型变量，通常假定是后者。

由式(3-25)两边求数学期望得到

$$E(z_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots\cdots + \beta_m x_{im} \quad \forall i \tag{3-26}$$

即 z_i 的数学期望 $E(z_i)$ 是变量 $x_{i1}, x_{i2}, \cdots\cdots, x_{im}$ 的线性函数。

二、回归系数的计算方法

假定对这些变量进行了 n 次观测，取得了 n 个观测样品，而且第 i 次观测得到的数据为 $(x_{i1}, x_{i2}, \cdots\cdots, x_{im}; z_i)$，$i = 1, 2, \cdots\cdots, n$，根据实际观测数据求回归模型(3-25)中的系数 $\beta_0, \beta_1, \beta_2, \cdots, \beta_m$ 的估计值 $b_0, b_1, b_2, \cdots, b_m$，即求回归方程

$$\hat{z} = b_0 + b_1 x_1 + b_2 x_2 + \cdots\cdots + b_m x_m \tag{3-27}$$

使得 n 个回归值 $\hat{z}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots + b_m x_{im}$ $(i = 1, 2, \cdots, n)$ 与 z 的观测值 z_i 的偏差平方和最小，亦即求偏回归系数 $b_0, b_1, b_2, \cdots, b_m$ 使函数

$$Q = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 = \sum_{i=1}^{n} (z_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im})^2 \tag{3-28}$$

达到最小值，于是偏回归系数 $b_0, b_1, b_2, \cdots, b_m$ 应满足如下方程

$$\frac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n} (z_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im}) = 0 \tag{3-29a}$$

$$\frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n}(z_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im})x_{i1} = 0 \tag{3-29b}$$

$$\vdots$$

$$\frac{\partial Q}{\partial b_j} = -2\sum_{i=1}^{n}(z_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im})x_{ij} = 0 \tag{3-29c}$$

$$\vdots$$

$$\frac{\partial Q}{\partial b_m} = -2\sum_{i=1}^{n}(z_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im})x_{im} = 0 \tag{3-29d}$$

将上述方程整理,可得到关于偏回归系数 $b_0, b_1, b_2, \cdots, b_m$ 的正则方程

$$\sum_{i=1}^{n}(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots + b_m x_{im}) = \sum_{i=1}^{n} z_i \tag{3-30a}$$

$$\sum_{i=1}^{n}(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots + b_m x_{im})x_{i1} = \sum_{i=1}^{n} z_i x_{i1} \tag{3-30b}$$

$$\vdots$$

$$\sum_{i=1}^{n}(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots + b_m x_{im})x_{ij} = \sum_{i=1}^{n} z_i x_{ij} \tag{3-30c}$$

$$\vdots$$

$$\sum_{i=1}^{n}(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots + b_m x_{im})x_{im} = \sum_{i=1}^{n} z_i x_{im} \tag{3-30d}$$

若令 $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}$, $B = [b_0\ b_1 \cdots\ b_m]^T$, $Z = [z_1\ z_2 \cdots\ z_n]^T$,则方程组(3 - 30a) ~(3 - 30d)可表示成矩阵形式

$$(X^T X)B = X^T Z \tag{3-31}$$

式中,矩阵 X 是一个 $n \times (m+1)$ 矩阵,矩阵中的 x_{ij} 表示第 i 个样品在第 j 个变量上的观测值($i = 1,2,\cdots n; j = 1,2,\cdots m$)。方程(3 - 31)的右端矩阵是自变量与因变量的组合。

在多元线性回归中,当观测数据远多于变量数即 $n \gg m$ 时,矩阵 X 一般是列满秩的,从而可以解出

$$B = (X^T X)^{-1} X^T Z \tag{3-32}$$

在实际编程和应用中,常可对上述正则方程组按如下形式进行改造,令 $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} (j = 1,2,\cdots,m)$,它们分别代表因变量和自变量的平均值,再引入如下记号

$$l_{jj'} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}), (j, j' = 1,2,\cdots,m) \tag{3-33a}$$

$$l_{jz} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(z_i - \bar{z}), (j = 1,2,\cdots,m) \tag{3-33b}$$

对式(3 - 30a)取平均可得到如下关系

$$b_0 = \bar{z} - \bar{x}_1 b_1 - \bar{x}_2 b_2 \cdots - \bar{x}_m b_m \tag{3-34}$$

将式(3 - 34)代入式(3 - 28)有

$$
\begin{aligned}
Q &= \sum_{i=1}^{n}(z_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im})^2 \\
&= \sum_{i=1}^{n}(z_i - \bar{z} + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \cdots\cdots + b_m \bar{x}_m - b_1 x_{i1} - b_2 x_{i2} - \cdots\cdots - b_m x_{im})^2 \\
&= \sum_{i=1}^{n}[(z_i - \bar{z}) - b_1(x_{i1} - \bar{x}_1) - b_2(x_{i2} - \bar{x}_2)\cdots - b_m(x_{im} - \bar{x}_m)]^2
\end{aligned} \tag{3-35}
$$

残差 Q 相对于偏回归系数 b_j 求导得到

$$
\begin{aligned}
\frac{\partial Q}{\partial b_j} &= -2\sum_{i=1}^{n}[(z_i - \bar{z}) - b_1(x_{i1} - \bar{x}_1) - b_2(x_{i2} - \bar{x}_2)\cdots - b_m(x_{im} - \bar{x}_m)](x_{ij} - \bar{x}_j) \\
&= -2(l_{jz} - l_{j1}b_1 - l_{j2}b_2\cdots - l_{jm}b_m), j = 1,2,\cdots m
\end{aligned} \tag{3-36}
$$

令$\frac{\partial Q}{\partial b_j} = 0(j = 1,2,\cdots m)$，得到偏回归系数 $b_0, b_1, b_2, \cdots, b_m$ 的另一种形式的正则方程组

$$
\begin{cases}
l_{11}b_1 + l_{12}b_2 + \cdots + l_{1m}b_m = l_{1z} \\
l_{21}b_1 + l_{22}b_2 + \cdots + l_{2m}b_m = l_{2z} \\
\cdots\cdots\cdots\cdots\cdots\cdots \\
l_{j1}b_1 + l_{j2}b_2 + \cdots + l_{jm}b_m = l_{jz} \\
\cdots\cdots\cdots\cdots\cdots\cdots \\
l_{m1}b_1 + l_{m2}b_2 + \cdots + l_{mm}b_m = l_{mz}
\end{cases} \tag{3-37}
$$

变量 x_j 和 $x_{j'}$ 之间的相关系数为

$$
r_{jj'} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ij'} - \bar{x}_{j'})^2}} = \frac{l_{jj'}}{\sqrt{l_{jj}l_{j'j'}}} \tag{3-38}
$$

$$
(j = 1,2,\cdots,m; j' = 1,2,\cdots,m)
$$

在方程组(3－37)的第 j 个关系式两边同除以 $\sqrt{l_{jj}l_{zz}}$，得到用相关系数表示的第三种形式的正则方程组，即所谓的标准化正则方程组

$$
\frac{l_{j1}}{\sqrt{l_{jj}}}\frac{b_1}{\sqrt{l_{zz}}} + \frac{l_{j2}}{\sqrt{l_{jj}}}\frac{b_2}{\sqrt{l_{zz}}} + \cdots + \frac{l_{jm}}{\sqrt{l_{jj}}}\frac{b_m}{\sqrt{l_{zz}}} = \frac{l_{jz}}{\sqrt{l_{jj}l_{zz}}} \quad (j = 1,2,\cdots,m) \tag{3-39}
$$

对式(3－39)稍作变换即可得到

$$
\frac{l_{j1}}{\sqrt{l_{jj}l_{11}}}\frac{\sqrt{l_{11}}b_1}{\sqrt{l_{zz}}} + \frac{l_{j2}}{\sqrt{l_{jj}l_{22}}}\frac{\sqrt{l_{22}}b_2}{\sqrt{l_{zz}}} + \cdots + \frac{l_{jm}}{\sqrt{l_{jj}l_{mm}}}\frac{\sqrt{l_{mm}}b_m}{\sqrt{l_{zz}}} = \frac{l_{jz}}{\sqrt{l_{jj}l_{zz}}} \quad (j = 1,2,\cdots,m) \tag{3-40}
$$

利用相关系数的定义并记 $\tilde{b}_j = \frac{\sqrt{l_{jj}}}{\sqrt{l_{zz}}}b_j$ $(j = 1,2,\cdots,m, \tilde{b}_j$ 为标准化回归系数)，方程(3－40)可表示为

$$
r_{j1}\tilde{b}_1 + r_{j2}\tilde{b}_2 + \cdots + r_{jm}\tilde{b}_m = r_{jz}(j = 1,2,\cdots,m) \tag{3-41}
$$

方程(3－39)的标准化正则方程组可改写为

$$
\begin{cases}
r_{11}\tilde{b}_1 + r_{12}\tilde{b}_2 + \cdots + r_{1m}\tilde{b}_m = r_{1z} \\
r_{21}\tilde{b}_1 + r_{22}\tilde{b}_2 + \cdots + r_{2m}\tilde{b}_m = r_{2z} \\
\cdots\cdots\cdots\cdots\cdots\cdots \\
r_{m1}\tilde{b}_1 + r_{m2}\tilde{b}_2 + \cdots + r_{mm}\tilde{b}_m = r_{mz}
\end{cases} \tag{3-42}
$$

在用计算机求解时，常先求出标准化回归系数，再求回归系数。

三、显著性检验

1. m 个自变量综合效果检验

在实际应用中，往往事先无法知道自变量和因变量是否具有线性关系。因此，由上述计算得到的回归方程是否具有代表性或显著性，需要用相应的方法进行检验。这里我们选用离差平方的分析考察回归效果的显著性，为此，首先对因变量的离差平方和进行因子分解

$$
\begin{aligned}
l_{zz} &= \sum_{i=1}^{n}(z_i - \bar{z})^2 = \sum_{i=1}^{n}(z_i - \hat{z}_i + \hat{z}_i - \bar{z})^2 \\
&= \sum_{i=1}^{n}[(z_i - \hat{z}_i)^2 + (\hat{z}_i - \bar{z})^2 + 2(z_i - \hat{z}_i)(\hat{z}_i - \bar{z})] \\
&= \sum_{i=1}^{n}(z_i - \hat{z}_i)^2 + \sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2 = Q + U \qquad (3-43)
\end{aligned}
$$

这里用到了关系式 $\sum_{i=1}^{n}(z_i - \hat{z}_i)(\hat{z}_i - \bar{z}) = 0$，请读者自行验证。在式(3－43)中，$Q = \sum_{i=1}^{n}(z_i - \hat{z}_i)^2$ 是残差平方和，反映了随机因素对 z 的影响；$U = \sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2$ 是回归平方和，反映了 m 个自变量对因变量 z 的影响。可以证明，在式(3－43)中 l_{zz}, U, Q 三项的自由度分别为

$$f_z = n - 1, f_U = m, f_Q = n - m - 1 \qquad (3-44)$$

假设 $H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_m = 0$ 为真时，统计量

$$F = \frac{U/m}{Q/(n - m - 1)} = \frac{(n - m - 1)U}{mQ} \qquad (3-45)$$

服从自由度为$(m, n - m - 1)$的 F 分布。

在显著性水平 α 下可查 F 分布表(附表 4)得到临界值 $F_{\alpha/2}(m, n - m - 1)$和 $F_{1-\alpha/2}(m, n - m - 1)$。当实际计算得到的 $F > F_{\alpha/2}(m, n - m - 1)$或 $F < F_{1-\alpha/2}(m, n - m - 1)$，则否定原假设，认为回归方程是显著的；当 $F < F_{\alpha/2}(m, n - m - 1)$或 $F > F_{1-\alpha/2}(m, n - m - 1)$，接受原假设，认为回归模型中的系数都为 0，尽管计算得到的回归系数可能与 0 有较大差别，仍认为回归方程式是不显著的。

上述的 F 检验实际上是方差分析的内容之一，可用表 3－3 的方差分析表来进行。

表 3－3　方差分析表

方差来源	平方和	自由度	均　方	F
回归	$U = \sum_{j=1}^{m} b_j l_{jz}$	m	U/m	
剩余	$Q = l_{zz} - U$	$n - m - 1$	$Q/(n - m - 1)$	$\frac{U/m}{Q/(n - m - 1)}$
总和	$l_{zz} = \sum_{i=1}^{n}(z_i - \bar{z})^2$	$n - 1$		

2. 单变量显著性检验

以上是对 m 个变量的综合效果进行检验，我们还可以对某一变量 x_j 的重要性单独地进行检验，即检验如下假设

$$H_0 : \beta_j = 0 \tag{3-46}$$

为此，讨论变量 x_j 对回归平方和 U 的贡献，对于 m 个变量的回归方程，因变量的离差平方和的因子分解见式(3-43)。

若把变量 x_j 剔除，设对于 $m-1$ 个变量的回归方程有如下的分解式

$$l_{zz} = \hat{U} + \hat{Q} \tag{3-47}$$

将式(3-43)和式(3-47)相减得到

$$U - \hat{U} + Q - \hat{Q} = 0 \tag{3-48}$$

令 $V_j = U - \hat{U} = \hat{Q} - Q$，它代表由于把变量 x_j 剔除所引起的剩余平方和的增量或回归平方和的减量。它恒为非负值，而且数值愈大表明 x_j 愈重要，称为 x_j 的偏回归平方和。可以证明，当原假设 $H_0 : \beta_j = 0$ 为真时

$$F_j = \frac{V_j}{Q/(n-m-1)} \tag{3-49}$$

服从自由度为$(1, n-m-1)$的 F 分布。于是我们可以在一定的显著性水平下检验 x_j 的显著性程度。检验原理与前面对 m 个变量综合效果检验相似，在此不再赘述。

四、复相关系数和偏相关系数

1. 复相关系数

考虑回归平方和 U 在因变量的总离差 l_{zz} 中所占的比例 $N = \frac{U}{l_{zz}}$，用它来度量 z 与自变量 $x_1, x_2, \cdots, x_m$ 之间关系的密切程度。与两个变量间的简单依赖关系相似，称 N 的平方根

$$r = \sqrt{\frac{U}{l_{zz}}} = \sqrt{\frac{l_{zz} - Q}{l_{zz}}} = \sqrt{1 - \frac{Q}{l_{zz}}} \tag{3-50}$$

为因变量 z 与自变量 $x_1, x_2, \cdots, x_m$ 之间的复相关系数，它反映了 z 与自变量 $x_1, x_2, \cdots, x_m$ 之间的依赖程度。与简单的相关系数不同的是，这个相关系数的变化范围为 $0 \leqslant r \leqslant 1$，它不能取负值。可以证明它等于 z 与它的回归估计值 $\hat{z}$ 的简单相关系数。回归估计值 $\hat{z}$ 是自变量 $x_1, x_2, \cdots, x_m$ 的线性组合，而且是相关系数最大的线性组合，请读者自行验证这一结论。

2. 偏相关系数

在多变量系统中，变量之间的关系是很复杂的，任意两个变量都有可能存在相关关系，这些关系又相互影响。在这种情况下，两个变量的简单相关关系往往不能反映他们之间的真实关系，很可能出现一些假象。基于这种情况，为了分别研究某两个变量的相关关系，必须在消除其它变量的影响后计算相关系数，这就是求偏相关系数。设有 m 个自变量 $x_1, x_2, \cdots, x_m$，为了研究变量 x_1, x_2 之间的关系，而消除 $x_3, \cdots, x_m$ 的影响，以 $x_3, \cdots, x_m$ 为自变量，分别作以 x_1, x_2 为因变量的回归分析，设回归值为 $\hat{x}_1$ 和 $\hat{x}_2$，剩余为 $e_1 = x_1 - \hat{x}_1$ 和 $e_2 = x_2 - \hat{x}_2$，我们把这两个剩余值的简单相关系数定义为 x_1, x_2 的偏相关系数，记为 r_{x_1, x_2}，类似地也可以定义其它偏相关系数。

上面详细讨论了多元线性回归分析的基本原理、实现步骤和回归方程检验方面的内容，在实际工作中，并不是每个变量都对回归方程是显著的。我们总是希望最终求取的回归方程中既没有漏掉对因变量有显著作用的自变量，也不包含对因变量无显著作用的自变量。这就涉

及如何选择变量的问题。逐步回归即是解决如何挑选变量并确定回归方程的一种统计方法。限于篇幅，关于逐步回归的内容，请读者参考相关文献。

五、多元线性回归实例

【例 3-2】 已知河北省某地在 1969 年至 1970 年间，每月统计了地下水的平均水位的高度 $z(m)$、月开采量 $x_1(\times 10^4 t/m)$ 和上一月地下水位的高度 $x_2(m)$，测得的原始数据列于表 3-4 中，试用多元回归分析求 $z(m)$ 与 $x_1(\times 10^4 t/m)$ 和 $x_2(m)$ 的回归关系。

表 3-4　测得的 z 和 x_1, x_2 原始数据

序号	取样日期	$x_1(\times 10^4 t/m)$	$x_2(m)$	$z(m)$
1	1969.09	99.55	12.5	12.3
2	1969.10	90.68	12.3	12.3
3	1969.11	42.41	12.3	12.4
4	1969.12	21.51	12.4	12.6
5	1970.01	36.51	12.6	13.3
6	1970.02	70.70	13.3	12.6
7	1970.03	310.92	12.6	10.4
8	1970.04	564.37	10.4	9.3
9	1970.05	489.66	9.3	8.7
10	1970.06	463.45	8.7	8.0
11	1970.07	375.00	8.0	7.2
12	1970.08	127.42	7.2	9.2

解：设回归方程为 $\hat{z} = b_0 + b_1 x_1 + b_2 x_2$，由公式(3-34)和公式(3-37)得到

$$\begin{cases} l_{11} b_1 + l_{12} b_2 = l_{1z} \\ l_{21} b_1 + l_{22} b_2 = l_{2z} \end{cases} \text{及}$$

$$b_0 = \bar{z} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

式中，$l_{11} = \sum_{i=1}^{12} x_{i1}^2 - \frac{1}{n}\left(\sum_{i=1}^{12} x_{i1}\right)^2$；$l_{12} = l_{21} = \sum_{i=1}^{12} x_{i1} x_{i2} - \frac{1}{n}\left(\sum_{i=1}^{12} x_{i1}\right)\left(\sum_{i=1}^{12} x_{i2}\right)$；

$l_{22} = \sum_{i=1}^{12} x_{i2}^2 - \frac{1}{n}\left(\sum_{i=1}^{12} x_{i2}\right)^2$，$l_{1z} = \sum_{i=1}^{12} x_{i1} z_i - \frac{1}{n}\left(\sum_{i=1}^{12} x_{i1}\right)\left(\sum_{i=1}^{12} z_i\right)$；

$l_{2z} = \sum_{i=1}^{12} x_{i2} z_i - \frac{1}{n}\left(\sum_{i=1}^{12} x_{i2}\right)\left(\sum_{i=1}^{12} z_i\right)$

将上表数据代入上述公式并求解正则方程组得到

$b_1 = -0.052, b_2 = 0.603, b_0 = 5.25$

为了进行显著性检验，再求出

$l_{zz} = 49.83, U = b_1 l_{1z} + b_2 l_{2z} = 47.03, Q = l_{zz} - U = 2.80$

$f_z = 12 - 1 = 11, f_U = 2, f_Q = 11 - 2 = 9$

得到检验统计量 $F = 75.51$，复相关系数 $r = \sqrt{\frac{47.03}{49.83}} = 0.971$，估计标准误差为 $s = \sqrt{\frac{Q}{f_Q}} =$

$\sqrt{\frac{2.8}{9}}=0.558$，查得 $F_{0.005}(2,9)=10.11$（见附表4），因为 $F>F_{0.005}(2,9)$，拒绝原假设，所以在 $\alpha=0.01$ 的显著性水平下回归方程是显著的。

由得到的回归方程可知，每个月的地下水位平均高度与月开采量呈负相关关系，而与上一个月的地下水平均水位呈正相关关系。

◆ 思考和练习题 ◆

1. 什么是相关？相关和回归有什么异同点？

2. 用例3－1中表3－2的数据做出散点图，用MATLAB软件验证例子中回归方程，并对回归系数进行显著性检验。

3. 解释复相关系数、偏相关系数的含义，及它们与简单相关系数的区别。

4. 应用线性回归模型进行预测和估计需注意哪些问题。

5. 某油田的油气产量和单位成本资料如下表：

某油田6个月内原油产量与单位成本统计数据表

月　份	1	2	3	4	5	6
原油产量，吨	4	6	8	7	8	9
单位成本，元/吨	73	72	71	72	70	69

试做出单位成本对产量的一元线性回归模型。对模型的优度进行评价。对该模型的回归系数进行检验和模型整体进行 F 检验，计算该回归模型的估计误差。在95%的置信水平下估计产量为10吨时单位成本的置信区间。

第四章 多元统计分析及其在油气勘探开发中的应用

本章学习目的和要求：了解多元统计方法的基本概念和方法原理，深刻理解各种多元统计方法产生的实际背景，掌握多元统计方法的应用条件和范围。第二和第三章的知识是本章学习的基础。

多元统计分析简称多元分析，是统计学的一个重要分支。随着计算机应用技术的普及，多元统计分析的应用愈来愈广，已成为处理多维数据不可缺少的工具。多元统计分析是研究多个随机变量之间的相互依赖关系以及内在统计规律的一门统计学科。常用的多元统计方法有多元回归分析、趋势分析、聚类分析、因子分析、判别分析、对应分析、典型相关分析、路径分析和多维标度法等。由于回归分析已在第三章作了介绍，本章主要介绍多元统计分析中的趋势分析、因子分析、聚类分析和判别分析。

第一节 趋势面分析及其应用

一、概述

“趋势”是指事物发展总的趋向，它不受局部因素的影响而由总的规律所支配，如某一构造幅度或某构造层厚度的区域性变化，某一地化指标分布的总体规律等。趋势性变化广泛存在于沉积学、构造地质学、地层学、地球化学、石油和天然气地质学等学科所研究的各种资料中。

所谓趋势分析就是要对数据中所包含的三部分信息，即区域性变化的信息、局部变化的信息和随机变化的信息进行分析，排除干扰因素，找出区域性变化趋势，突出局部异常。例如，在油气勘探和开发中，研究油气地质特征在时空中的分布趋势和规律，是油气地质研究的重要问题之一。

趋势分析是用一个适当的数学函数(一维是直线，二维是平面，三维是空间体)去拟合观测数据，把观测数据中受区域影响的在大范围内变化的特点与受局部影响的在小范围内变化的特点区分开。根据研究的问题不同，有时需要将局部影响因素分离出来，研究局部异常的变化特征，如油气物、化探异常的分析等。有时注重全局性的变化趋势，要将大范围内的变化特点提取出来，例如，对区域构造基底物探异常的研究。至于何时区分全局性的变化和局域性变化，要看研究问题的具体特点而定。

根据研究对象所处的空间维数不同，趋势分析有一维、二维和多维趋势分析。根据趋势模型中所含变量的次数，趋势分析可分为一次、二次和多次趋势分析。为了表示地质变量在空间或时间内的周期性变化，还可以用三角多项式构造趋势面函数，此时称为调和趋势分析。当然，根据具体问题的性质还可以选很多其它的趋势函数。由于多项式趋势函数和三角多项式趋势函数能在有限范围任意逼近各种连续函数，在实际工作中应用较广，下面以多项式趋势面分析为例介绍趋势分析的数学函数计算、趋势面拟合程度检验和应用实例。其他各种趋势分

析在原理上与之相同，都是用最小二乘法化成回归分析的问题。

二、多项式趋势面的计算

多项式趋势面分析所用的拟合函数是多项式，它具有计算方便和所代表的曲面易于观测的优点，可以拟合任意的复杂数据，在实际工作中已得到广泛应用。这里以二维趋势面分析为例，导出趋势面方程和介绍趋势分析的实现过程。

用(x,y)表示样点的平面坐标，函数$f(x,y)$表示实际观测值z在大范围内的变化趋势，该函数表示的曲面称为趋势面，则实际观测值z可表示成

$$z = f(x,y) + R(x,y) \tag{4-1}$$

式中，$R(x,y)$称为剩余，它由局部性变化和随机性变化引起。

若能从实际观测数据中把$f(x,y)$和$R(x,y)$计算出来，就能把全局性变化与局域性变化或随机性变化区分开，这就是趋势面分析的基本思想。实际工作中，通常选用某种特殊函数表示趋势面函数$f(x,y)$，最常见的是把$f(x,y)$表示成多项式，例如二次多项式可表示成

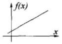
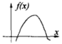
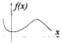
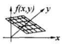
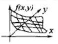
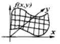
$$f(x,y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2 \tag{4-2}$$

此时，式(4-1)可写成如下形式

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i^2 + \beta_4 x_i y_i + \beta_5 y_i^2 + R_i \qquad \forall i \tag{4-3}$$

表4-1列出了常见的各种趋势函数的数学关系和示意图形。

表4-1　常见的各种趋势函数的数学关系和示意图形

分　类	次　数	数学方程式	系数	几何图形
一维	一次	$f(x)=\beta_0+\beta_1 x$	1	
	二次	$f(x)=\beta_0+\beta_1 x+\beta_2 x^2$	2	
	三次	$f(x)=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3$	3	
二维	一次	$f(x,y)=\beta_0+\beta_1 x+\beta_2 y$	2	
	二次	$f(x,y)=\beta_0+\beta_1 x+\beta_2 y+\beta_3 x^2 + \beta_4 xy+\beta_5 y^2$	5	
	三次	$f(x,y)=\beta_0+\beta_1 x+\beta_2 y+\beta_3 x^2+\phi_4 xy + \beta_5 y^2+\beta_6 x^3+\beta_7 x^2 y+\beta_8 xy^2 + \beta_9 y^3$	9	

设在观测区内有n个观测点，各点的坐标和z的观测值如表4-2所示，观测点变量的二次趋势值可以表示成

$$\hat{z}_i = b_0 + b_1 x_i + b_2 y_i + b_3 x_i^2 + b_4 x_i y_i + b_5 y_i^2 \quad (i = 1,2,\cdots,n) \tag{4-4a}$$

式中，$b_0, b_1, b_2, \cdots, b_5$是式(4-3)中各系数$\beta_0, \beta_1, \beta_2, \cdots, \beta_5$的估计值。

为了书写方便，设 $x_{i1}=x_i, x_{i2}=y_i, x_{i3}=x_i^2, x_{i4}=x_iy_i, x_{i5}=y_i^2$，则式(4-4a)变为

$$\hat{z}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} \quad (i=1,2,\cdots,n) \qquad (4-4b)$$

由实测数据，利用第三章介绍的多元线性回归得到回归系数 $b_0, b_1, b_2, \cdots, b_5$

$$B = (X^TX)^{-1}X^TZ \qquad (4-5)$$

其中，$B=[b_0 \quad b_1 \quad \cdots \quad b_5]^T, X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{15} \\ 1 & x_{21} & \cdots & x_{25} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n5} \end{bmatrix}, Z=[z_1 \quad z_2 \quad \cdots \quad z_n]^T$。

将由式(4-5)得到的回归系数代入式(4-4b)即可得到不同空间点(x_i, y_i)处的趋势值 $\hat{z}_i$。

对于一般的 l 次多项式，除常数项以外，共有 $m=l(l+3)/2$ 项，这时的趋势分析数学模型为

$$z_i = \beta_0 + \beta_1x_i + \beta_2y_i + \beta_3x_i^2 + \beta_4x_iy_i + \beta_5y_i^2 + \cdots\cdots + \beta_my^l + R_i \quad \forall i \qquad (4-6)$$

若把其中的每一项看作一个变量，不妨设

$$x_{i1}=x_i, x_{i2}=y_i, x_{i3}=x_i^2, x_{i4}=x_iy_i, x_{i5}=y_i^2, \cdots\cdots, x_{im}=y_i^l \qquad (4-7)$$

此时式(4-6)的趋势分析数学模型可改写为

$$z_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + \beta_4x_{i4} + \beta_5x_{i5} + \cdots\cdots + \beta_mx_{im} + R_i \quad \forall i \qquad (4-8)$$

表 4-2　二维趋势分析各点坐标和观测值

点　号	x	y	z	点　号	x	y	z
1	x_1	y_1	z_1	⋮	⋮	⋮	⋮
2	x_2	y_2	z_2	n	x_n	y_n	z_n

仿照上面二次趋势面分析的做法，用最小二乘法得到方程

$$\hat{z}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + \cdots\cdots + b_mx_{im} (i=1,2,\cdots,n) \qquad (4-9)$$

的系数为

$$B = (X^TX)^{-1}X^TZ \qquad (4-10)$$

把由式(4-10)计算得到的系数代入式(4-9)中，可得到观测点的趋势值 $\hat{z}_i(i=1,2,\cdots,n)$，则观测点的剩余为

$$R_i = z_i - \hat{z}_i \quad (i=1,2,\cdots,n) \qquad (4-11)$$

三、趋势面拟合程度的检验

一般来说，任何一组观测数据都可以用适当次数的多项式来表示其变化趋势。但这个多项式是否能较好地代表原始变量 Z 的空间变化趋势，这个趋势多项式对变量的拟合程度如何，需要用统计方法进行检验。统计检验的方法有很多，这里仅利用第二章介绍的统计参数的显著性检验方法。

在假设剩余 $R_i(i=1,2,\cdots,n)$相互独立的条件下，对因变量的离差平方和做如下变换

$$l_{zz} = \sum_{i=1}^{n}(z_i-\bar{z})^2 = \sum_{i=1}^{n}(\hat{z}_i-\bar{z})^2 + \sum_{i=1}^{n}(z_i-\hat{z}_i)^2 = U + Q \qquad (4-12)$$

其中 l_{zz}的自由度$f_{zz}=n-1$；U 的自由度$f_U=m$；Q 的自由度$f_Q=n-m-1$。

式(4-12)表明变量的总离差平方和 l_{zz} 可以分解成趋势值 $\hat{z}$ 的总离差平方和 U 与剩余值的平方和 Q。这样可以定义趋势面的拟合优度系数为

$$r^2 = \frac{U}{l_{zz}} = 1 - \frac{Q}{l_{zz}} \tag{4-13}$$

从式(4-13)可以看到,$0 \leqslant r^2 \leqslant 1$,它在一定程度上反映趋势面的拟合程度。但在实际工作中发现,它还与取样点的数目关系极大。从式(4-9)的线性方程组解的分析可以知道,当 $n = m+1$ 时,方程组满秩,这时趋势值与原始变量完全拟合且 $r^2 = 1$。当 n 略大于 $m+1$ 时,数据的拟合度接近 1。由此可见,拟合度 r^2 仅反映 $\hat{z}$ 对观测值平均 $\bar{z}$ 的拟合程度,并不能检验多项式的显著性。为此,一般要利用式(4-12)的离差平方和相关的量,并考虑自由度在内的 F 统计量对多项式进行显著性检验

$$F = \frac{U/m}{Q/(n-m-1)} = \frac{(n-m-1)U}{mQ} \tag{4-14}$$

这里所要检验的假设是

$$H_0: \beta_1 = \beta_2 = \cdots\cdots = \beta_m = 0$$

所谓检验结果不显著,即接受原来的假设 H_0,意味着多项式所有系数的理论值都为 0,只要有一个系数显著异于 0,则必须否定假设 H_0。由此可见,这种假设检验方法是很粗糙的,一般总会得到显著性的结论。

比较精细的方法是对趋势面的拟合度逐次进行检验。假定相继算出了$(l-1)$次和 l 次趋势多项式,它们的趋势值离差平方和及剩余平方和分别为 U_{l-1}, U_l, Q_{l-1} 和 Q_l,于是有

$$l_{zz} = U_{l-1} + Q_{l-1} = U_l + Q_l \tag{4-15}$$

当趋势多项式的次数升高时,趋势值的离差平方和不减,而剩余平方和也不增,即有

$$U_{l-1} \leqslant U_l, Q_{l-1} \geqslant Q_l \tag{4-16}$$

在 l 次多项式趋势分析中,预测关系式中含有$(l+1)$个 l 次项

$$x^l, x^{l-1}y, \cdots\cdots, xy^{l-1}, y^l$$

它们的系数分别为 $b_{m-l}, b_{m-l+1}, \cdots\cdots, b_{m-1}, b_m$,增量 $V_l = U_l - U_{l-1}$ 可以度量这些 l 次多项式对趋势拟合所做的贡献,以它为基础构造新的 F 统计量

$$F_l = \frac{V_l/(l+1)}{Q_l/(n-m-1)} = \frac{(n-m-1)V_l}{(l+1)Q_l} \tag{4-17}$$

实际应用中,需检验 l 次项系数的显著性,即假设检验

$$H_0: \beta_{m-l} = \beta_{m-l+1} = \cdots\cdots = \beta_{m-1} = \beta_m = 0$$

当 H_0 为真时,F_l 服从自由度为$(l+1, n-m-1)$的 F 分布。由 F 分布表(附表4)可按一定的显著性水平 α 查出相应的临界值 $F_{\alpha/2}(l+1, n-m-1)$ 和 $F_{1-\alpha/2}(l+1, n-m-1)$。若算出的 F_l 值大于 $F_{\alpha/2}(l+1, n-m-1)$ 或小于 $F_{1-\alpha/2}(l+1, n-m-1)$,则否定 H_0 假设,即认为这些系数中至少有一个显著地异于零,从$(l-1)$次多项式升高到 l 次多项式是有意义的,否则,接受 H_0,认为从$(l-1)$次多项式升高到 l 次多项式是没有意义的。

四、应用实例

【例 4-1】 有某油层 18 口井的钻井资料,其中某层的厚度变化如表 4-3 所示,求该层厚度变化的二次趋势面,并检验其显著性,作出其剩余分布图,指出正常场分布。

表 4-3 某层的厚度变化数据表

测点号	坐标		厚度 z,m	测点号	坐标		厚度 z,m
	x	y			z	y	
1	4.21	2.85	231.0	10	4.59	3.22	259.0
2	3.93	1.70	248.5	11	4.59	3.23	259.8
3	5.04	1.55	196.0	12	4.10	2.84	261.5
4	3.85	1.11	211.0	13	4.68	2.75	272.0
5	4.53	3.22	248.5	14	4.70	2.75	276.5
6	4.58	2.75	244.0	15	3.77	1.32	253.5
7	4.70	2.17	590.0	16	3.91	3.48	259.5
8	4.89	2.72	280.4	17	4.57	3.22	225.5
9	4.56	3.19	241.0	18	2.48	2.80	613.0

解:采用二次趋势面方程

$$\hat{z} = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 \tag{4-18}$$

(1)趋势面系数满足的方程组

$$\begin{bmatrix} 18.0 & 77.7 & 46.9 & 341.1 & 203.2 & 130.9 \\ & 341.1 & 203.2 & 1517.7 & 896.5 & 568.7 \\ & & 130.9 & 896.5 & 568.7 & 381.2 \\ & & & 6818.6 & 4001.7 & 2511.7 \\ & & & & 2511.7 & 1657.7 \\ & & & & & 1139.9 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 5170.7 \\ 21782.7 \\ 13503.5 \\ 94446.7 \\ 56969.8 \\ 37416.9 \end{bmatrix}$$

(2)解得系数

$b_0 = 1189.3$, $b_1 = -645.6$, $b_2 = 502.4$

$b_3 = 73.6$, $b_4 = -8.2$, $b_5 = -100.0$

(3)计算剩余的平方和

$U = 132081, f_U = 5$

$Q = 98910, f_Q = 18 - 5 - 1 = 12$

$l_{zz} = 230911, f_{zz} = 17$

(4)计算统计量

拟合度 $r^2 = \dfrac{132081}{230911} = 57.2\%$

统计量 $F = \dfrac{132081 \times 12}{98910 \times 5} = 3.2$

(5)显著性检验

若取显著性水平 $\alpha = 0.1$,查 F 分布表(附表 4),$F_{0.05}(5,12) = 3.11$,有 $F > F_{0.05}(5,12)$。因此,在显著性水平 $\alpha = 0.1$ 下,趋势面拟合是显著的。

图 4-1(a)和图 4-1(b)分别为厚度分布的二次趋势面和剩余等值线分布图。从图 4-1(b)可以看到在 B 点附近区域有明显的正异常。

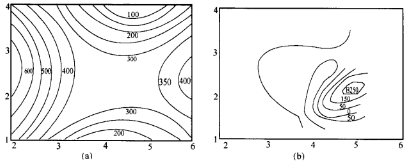


图 4-1　厚度分布的二次趋势面(a)和剩余等值线分布图(b)

五、趋势面剩余分析和其他问题讨论

从上面的分析可以看到,趋势面分析的思想看起来比较简单,但在应用中还有许多理论问题和实际问题没有取得一致的看法。例如,用趋势分析解决系统趋势和确定局部异常这两个性质不同的问题时,数学模型应有本质的差异,但由于应用的数学方法一样,或在同一个问题中同时期望达到两个目的,人们往往不太注意数学模型上的差异。在趋势面分析中,假设了有剩余 $R_i(i=1,2,\cdots,n)$存在,但在计算时并没有对剩余 $R_i(i=1,2,\cdots,n)$的性质进行分析。

事实上,针对系统趋势分析和确定局部异常这两个模型,对剩余 $R_i(i=1,2,\cdots,n)$应有不同的理解。在系统趋势分析中,剩余 $R_i(i=1,2,\cdots,n)$仅代表随机误差,通常假设对不同的观测点(x_i,y_i)剩余 R_i 相互独立且满足高斯正态分布

$$R_i = e_i,\quad e_i \sim N(0,\sigma^2)(i = 1,2,\cdots,n) \tag{4-19}$$

而在确定局部异常的模型中,剩余 R_i 中应包括局部异常 r_i 和随机误差 e_i 两部分,此时的局部异常 r_i 不是独立的,而假设随机误差 e_i 是独立的且满足高斯正态分布,即

$$R_i = r_i + e_i,\quad e_i \sim N(0,\sigma^2)(i = 1,2,\cdots,n) \tag{4-20}$$

在研究局部变化规律时,剩余值 R_i 中包含反映局部变化的信息。实际工作中发现,观测值 z_i 在趋势值$\hat{z}_i$ 和局部异常 r_i 两者之和的附近摆动。对于不同的观测点,随机误差 e_i 是相互独立的且满足正态分布,而局部异常 r_i 却不独立,从而剩余 R_i 也不独立。剩余值相关性较强的部位,往往指示有意义的异常区域,那么多大的剩余量才能分离出局部异常呢?针对这一实际问题,一般需要确定一个异常的下限值 $\delta>0$,当剩余量高于该异常下限值时,则可分离出局部异常,否则,剩余量仅代表随机变化。下面给出确定异常下限的一种经验方法。

若剩余值满足 $R_i=e_i\sim N(0,\sigma^2)(i=1,2,\cdots,n)$,则满足条件 $P\{-\delta\leqslant R_i\leqslant\delta\}=0.95$ 的 δ 为

$$\delta = 1.96\sigma$$

式中的 σ 是未知的,需用其估计值来代替,即 $\hat{\sigma}=\sqrt{Q/(n-m-1)}$,则异常的下限为

$$\delta = 1.96\hat{\sigma} = 1.96\sqrt{Q/(n-m-1)}$$

式中,$(n-m-1)$是剩余值的自由度,见表 3-3。

当然还可以取全部正剩余的平均值或作剩余的直方图,取其众数值做为异常的下限。

趋势面分析中的另一个问题是趋势面次数的确定。从数学原理上分析,随着趋势面次数的

升高,拟合效果会得到一定的改善。但在实际工作中并不是趋势面的次数愈高愈好。在系统趋势分析中,拟合程度愈高愈好,而在局部异常分析中,过度的拟合有可能使剩余异常变小,从而将一部分有意义的局部异常归到系统趋势中去,在这种情况下趋势面的拟合次数不宜太高。

最后的一个问题就是趋势面的畸变和正规方程组的病态性。正规方程的病态性是指由于个别测量数据的错误或误差造成趋势面方程式(4-9)的系数矩阵为奇异阵,得到趋势面方程的系数是一个奇异解的现象。在处理实际数据时常会碰到一些反常现象,如对正值变量做趋势分析会出现负值,或出现特高值或特低值,这种现象称为趋势面的畸变,尤其在没有控制点的区域易于出现。

第二节　因子分析及其应用

因子分析方法是由 Spearman 在 19 世纪初研究心理学问题时提出的,以后逐渐应用到社会与自然科学各个领域。它主要用于研究一组样本或一组变量之间的相关性,1957 年由 Krumbein 引进地质学。后来,Imbrie 对因子分析在地质学中的应用和发展作了大量工作。

因子分析主要研究能体现变量或样本数据之间的相关关系的相关阵的内部结构,将许多原始变量组合成少数几个"因子"。该方法可以帮助我们对大量地质观测资料进行分析,提取原始变量中的大部分相关信息与变异信息,对起决定作用的重要因素作出较为合理解释,是研究系统分类、成因分类的重要手段。通过因子分析,可以再现原始变量之间的相关关系,揭示产生这些关系的内在原因,有助于探索事物的因果关系。诸多地质现象的成因研究是地学研究的重要内容,因子分析可以从以下几个方面为成因研究提供帮助:

第一,压缩原始数据。在研究地质问题时,人们希望收集尽可能多的资料或数据,而在综合分析这些数据以形成成因概念时,又为面对众多的复杂、有时还是互相矛盾的数据而苦恼。因子分析为这些众多复杂数据的精简提供了一种数学算法,它能在数量上大大精简原始数据但又不损失数据中包含的成因信息。这样就有利于地质人员进行综合分析。

第二,指示成因推理方向。因子分析能够把庞杂纷乱的原始数据按成因上的联系进行归纳、整理、精炼和分类,理出几条客观的成因线索,为地质人员提供逻辑推理方向,启发思考相应的成因结论。

第三,分解叠加的地质过程。现实中观测到的地质现象往往是多种成因过程叠加的产物,因子分析提供了一个分解叠加过程而识别每个单一地质过程的手段。

鉴于上述原因,因子分析在地学领域的应用十分广泛,已有效地应用于沉积盆地蚀源区的研究、沉积物的粒度分析、沉积相研究、地层分析、古环境与古生态的研究、石油及天然气成因研究、油田水化学研究、有机地球化学研究及石油、天然气化探资料分析等各个方面。

因子分析以研究的目标是变量的相关关系还是样本的相关,可分为 R 型分析和 Q 型分析两种。立足于变量之间的相关性研究的叫做 R 型因子分析,立足于样品之间的相关性研究的称为 Q 型因子分析。现以主成分分析作为 R 型因子分析的起点讲述因子分析的方法原理。

一、主成分分析

1.主成分分析及基本思想

1)什么是主成分分析

主成分的概念最早由数学家 *Karl Parson* 在 1901 年提出的,当时只是针对非随机变量来讨

论的，1933 年 *Hotelling* 将这个概念推广到随机变量。1957 年 *Krumbein* 把该方法引入到地质学研究，后来，又由 *Imbrie* 对该方法在地质学中的应用和发展作了大量的基础工作。

主成分分析是把数目较多的变量作线性组合，组合成几个能反映主要地质特征的新变量，将原有的多种指标重新组合成一组新的互不相关的几个综合指标来代替原来的指标，同时根据实际需要从中取较少几个综合指标，而且被选取的几个指标要尽可能多地反映原来指标的信息。在地质学研究中，常常要作多变量的综合分析，这些变量经常不是独立的，存在复杂的相关关系，使综合分析出现困难，为使分析层次更分明，需要用到主成分分析，即数学上处理降维的一种方法。

日常生活中主成分分析的思想也常用到。例如，要做某人的一件衣服，需测量许多尺寸，如身长、袖长、胸围、腰围、肩宽和肩厚等十几项指标，但某一服装厂要生产一批服装不可能针对具体的某个人把尺寸分得过多，而只能把多个指标综合成几个综合指标，作为分类的型号。利用主成分分析将上述十几项指标综合成三项指标，一项是反映长度的指标，一项是反映胖瘦的指标，一项是反映特体的指标。

主成分分析还可以和其它方法结合使用，如和回归分析结合便形成主成分回归，它可以克服回归分析中由于自变量之间的高度相关而产生的问题。

2)主成分分析的基本思想

主成分分析在数学上就是将原来 m 个指标作线性组合，求得新的综合指标，并选取几个具有代表性的综合指标(原指标的线性组合)。下面介绍这种选择的方法原理和实现过程。

如果将选取的第一个线性组合即第一个综合指标记为 F_1，自然希望 F_1 尽可能多地反映原来的指标信息，这里的"信息"用什么来表示呢？最经典的方法就是用 F_1 的方差来表示，F_1 的方差 $Var(F_1)$ 愈大，表示 F_1 包含的信息愈多。因此，在所有的线性组合中，选取的 F_1 应该是方差最大的，称 F_1 为第一主成分。如果第一主成分 F_1 没有包含原来 m 个指标的绝大部分信息，须考虑选取第二个线性组合 F_2，且希望 F_1 中已有的信息不出现在 F_2 中，即：

$$Cov(F_1, F_2) = 0 \tag{4-21}$$

F_2 称为第二主成分，依此可以建立第三、第四等主成分，要求这些主成分互不相关，且方差依次减小。

2. 主成分分析的几何意义和数学模型

下面用一个简单的例子从几何上说明主成分分析的原理。设有 n 个样品，每个样品测量了两个变量 x_1 和 x_2，在由 x_1 和 x_2 确定的样品空间中，n 个样品点的分布如图 4-2 所示。

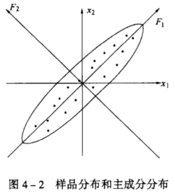


图 4-2　样品分布和主成分分布的关系

从图 4-2 可以看到，变量 x_1 和 x_2 都有较大的波动(方差较大)，而且，二者具有明显的相关性。如果作一坐标旋转，取 F_1 和 F_2 为新坐标轴，则在新坐标系中有如下性质：

(1) n 个样品点的新坐标 F_1 和 F_2 的相关性很小，几乎为 0；

(2)在新坐标 F_1 和 F_2 中，n 个点的波动(方差)，大部分归结为 F_1 的波动，而 F_2 的波动很小，故用 F_1 可以反映变量的大部分信息；

(3) F_1 和 F_2 与 x_1 和 x_2 间的关系可用下式表示：

$$F_1 = x_1\cos\theta + x_2\sin\theta$$
$$F_2 = x_2\cos\theta - x_1\sin\theta \qquad (4-22)$$

式(4-31)的矩阵表达形式为:

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad (4-23)$$

其中 $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ 是标准正交矩阵,其元素满足:

$$\begin{cases} a_{j1}^2 + a_{j2}^2 = 1 \\ a_{11}a_{21} + a_{12}a_{22} = 0 \end{cases} \quad (j = 1,2) \qquad (4-24)$$

从以上三点可以看出,这样得到 F_1 的 F_2 和是无关的,而且 F_1 在 x_1 和 x_2 的线性组合中方差最大,线性组合的系数 $a_{jj'}(j=1,2;j'=1,2)$ 满足标准正交的条件。

将以上结果推广到 m 维的情况,设有 n 个样品,每个样品有 m 个变量,经过适当的线性组合,可得到 m 个新变量:

$$F_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jm}x_m (j = 1,2,\cdots,m) \qquad (4-25)$$

此时,矩阵 $A = [a_{jj'}]_{m\times m}$ 中的系数 $a_{jj'}(j,j'=1,2,\cdots,m)$ 满足如下条件:

(1) F_j, $F_{j'}$ 是线性无关的 $(j \neq j')$;

(2)线性组合的系数 $\sum_{j'=1}^{m} a_{jj'}^2 = 1(j = 1,2,\cdots,m)$;

(3) F_1 在 $x_1,x_2,\cdots,x_m$ 的所有线性组合中方差最大;F_2 与 F_1 不相关且除 F_1 外在 $x_1,x_2,\cdots,x_m$ 的所有线性组合中方差最大;F_j 与 $F_1,F_2,\cdots,F_{j-1}$ 不相关且除 $F_1,F_2,\cdots,F_{j-1}$ 外在 $x_1,x_2,\cdots,x_m$ 的所有线性组合中方差最大。

这样确定的新变量 $F_1,F_2,\cdots,F_{j-1},\cdots,F_m$ 称为原始变量的第1、第2、…、第 m 个主成分,其中 F_1 在总的方差中占的比例最大,最重要,其余的 $F_2,\cdots,F_{j-1},\cdots,F_m$ 的方差依次递减,其重要性也依次减小,这样就可以取前面少数几个主成分对样本数据的主要性质进行分析。

3.主成分的计算

确定主成分的问题,在几何上就是要找出 m 维空间中样本集的投影主轴,在代数上就是找出 m 个变量的协方差矩阵的特征向量。下面就代数方法计算主成分的原理和实现步骤开展讨论。

第一步:原始数据的标准化

设有两个随机变量 x_j 和 $x_{j'}$,n 个观测值为:

$$x_j = (x_{1j},x_{2j},\cdots,x_{nj})^T, x_{j'} = (x_{1j'},x_{2j'},\cdots,x_{nj'})^T \qquad (4-26)$$

上述变量称为原始随机变量,由于变量的量纲和数值的差别,在作主成分分析时,需对变量进行标准化处理,即:

(1)计算各变量的均值: $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $\bar{x}_{j'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij'}$

(2)求出各变量的方差: $\sigma_{x_j}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, $\sigma_{x_{j'}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij'} - \bar{x}_{j'})^2$

(3)变量标准化：令新变量 $x'_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_{x_j}}, x'_{ij'} = \frac{x_{ij'} - \overline{x}_{j'}}{\sigma_{x_{j'}}}, i = 1,2,\cdots n$，标准化后的变量满足条件：

$$\overline{x'}_j = \frac{1}{n}\sum_{i=1}^{n} x'_{ij} = 0, \overline{x'}_{j'} = \frac{1}{n}\sum_{i=1}^{n} x'_{ij'} = 0 \tag{4-27}$$

$$\sigma^2_{x'_j} = \frac{1}{n}\sum_{i=1}^{n} x'^2_{ij} = 1, \sigma^2_{x'_{j'}} = \frac{1}{n}\sum_{i=1}^{n} x'^2_{ij'} = 1 \tag{4-28}$$

为了书写方便，后面仍用 x_j 和 $x_{j'}$ 代表标准化处理后的变量，即假定 x_j 和 $x_{j'}$ 是标准化变量，它们的方差和相关系数为：

$$\sigma^2_{x_j} = \frac{1}{n}\sum_{i=1}^{n} x^2_{ij} = \frac{1}{n}x_j^T x_j = 1, \sigma^2_{x_{j'}} = \frac{1}{n}\sum_{i=1}^{n} x^2_{ij'} = \frac{1}{n}x_{j'}^T x_{j'} = 1 \tag{4-29}$$

$$r^2_{x_j x_{j'}} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'} = \frac{1}{n}x_j^T x_{j'} \tag{4-30}$$

由上述公式可以看出，对标准化变量 x_j，它的长度为：

$$\| x_j \| = \sqrt{x_j^T x_j} = \sqrt{n} \tag{4-31}$$

两个不相关的标准化变量 x_j 与 $x_{j'}$ 的相关系数为0，从代数上说相当于它们的内积 $x_j^T x_{j'} = 0$，从几何上说相当于两个向量直交。

第二步：相关矩阵计算

从所研究的地质对象中，取 n 个样品，对每个样品测得 m 个变量的值，得原始数据阵：

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} = [x_1, x_2, \cdots, x_j, \cdots, x_m] \tag{4-32}$$

式中的每一行代表一个观测点，每一列代表一个变量，x_{ij} 代表第 i 个样品在第 j 个变量上的观测值，$x_j = (x_{1j}, x_{2j} \cdots x_{nj})^T, (j = 1,2,\cdots,m)$，代表第 j 个变量的 n 个观测值。

上述原始数据阵的每个列，可以看成是 m 维欧氏空间的一个向量，用 m 个向量 $x_j (j = 1, 2,\cdots m)$ 代替 m 个随机变量，通过研究这 m 个向量的空间位置关系来研究原来 m 个随机变量之间的关系。m 个向量两两之间的相关系数如公式(4－39)所示，组成一个 m 阶方阵 R 为：

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} = \frac{1}{m}X^T X \tag{4-33}$$

这个 m 阶方阵 R 叫做这 m 个变量的相关阵，它是半正定对称的，因此其特征值都是非负的。若 m 个变量是线性无关的，则 R 是正定的，其特征值为正值。相关阵是主成分分析方法的基础，研究相关阵的结构也是因子分析的重要内容。

第三步：特征值与主成分得分的计算

(1)求出相关矩阵 R 的特征值:

设 $U=(u_1,u_2,\cdots,u_m)^T$ 为相关矩阵 R 的特征向量矩阵,$\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_m)$为相关矩阵 R 的特征值矩阵,则有

$$RU = \lambda U \tag{4-34}$$

求解 R 的特征值$\lambda_j(j=1,2,\cdots m)$并按大小排列为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$,并求对应于特征值 λ_j 的特征向量 u_j。

(2)确定主成分的个数 p:

一般按 p 个特征值之和占所有特征值之和的百分数大于 90%来确定应取的主成分个数 p,并计算主成分 $F_j(j=1,2,\cdots,p)$:

$$F_j = u_{1j}x_1 + u_{2j}x_2 + \cdots + u_{pj}x_p \tag{4-35}$$

(3)计算各样品在 p 个主成分上的得分,即对每个样品计算出它的前 p 个主成分,第 i 个样品的第j 个主成分的得分为:

$$F_{ij} = u_{1j}x_{i1} + u_{2j}x_{i2} + \cdots + u_{pj}x_{ip}(i = 1,2,\cdots n; j = 1,2,\cdots p) \tag{4-36}$$

(4)利用 p 个主成分作地质解释或利用样品在每个主成分上的得分进行分类。

二、R 型因子分析

将主成分分析向前推进一步就可以得到 R 型因子分析。R 型因子分析中 m 个变量间的相互关系是通过p 个公因子来表示的。与前述主成分的表达式(p 个主成分通过变量来表示)比较,因子分析采用的模型其统计意义更明确,有助于地质解释。

1.因子模型

设有 n 个样品,每个样品观测了 m 个变量。将原始数据标准化后,得标准化数据阵如下

$$X = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_m \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \end{array} = [X_1 \quad X_2 \quad \cdots \quad X_n] \tag{4-37}$$

各变量之间的相关系数组成的相关矩阵为

$$R = \frac{1}{n-1}XX^T$$

求出 R 的特征值$\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_m\geqslant 0$ 及对应的单位特征向量 $u_1,u_2,\cdots,u_m$ 组成的正交矩阵 $U=[u_1,u_2,\cdots,u_m]$。此时如果用向量 $F_i=(F_{i1},F_{i2},\cdots,F_{im})^T,(i=1,2,\cdots,n)$表示第 i 个样品的 m 个主成分,那么根据上一部分主成分分析的内容,n 个样品的主成分数据阵可表示为特征向量矩阵 $U=[u_1,u_2,\cdots,u_m]$与原始数据阵的乘积

$$\begin{aligned} F &= (F_1,F_2,\cdots,F_n) = U^TX = U^T[X_1,X_2,\cdots,X_n] \\ &= [U^TX_1,U^TX_2,\cdots,U^TX_n] \end{aligned} \tag{4-38}$$

实际上,我们不需要 m 个主成分,只需要其中一部分就足以代表 m 个变量的变化,通常选取 p 个主成分,使

$$\frac{\sum_{j=1}^{p}\lambda_j}{\sum_{j=1}^{m}\lambda_j} > 90\% \tag{4-39}$$

它表示前 p 个主成分已解释了全部变量变化的 90%以上的信息。这 p 个主要成分对应的数据矩阵就是将特征向量矩阵 $U=[u_1,u_2,\cdots,u_m]$剖分为

$$U = [\underbrace{u_1,\cdots,u_p}_{p}\ ,\underbrace{u_{p+1},\cdots,u_p}_{m-p}] = [\underset{m\times p}{U_1}\ \underset{m\times(m-p)}{U_2}] \tag{4-40}$$

的前面的 p 个向量组成的矩阵。将式(4-40)代入式(4-38),并将 U_1 与 U_2 乘积对应的部分分开

$$\underset{m\times N}{F_{(1)}} = \underset{m\times p}{U_1^T}\ \underset{p\times N}{X} \tag{4-41}$$

由于 $F=U^TX$,故有

$$X = UF = [U_1 \quad U_2]\begin{bmatrix} F_{(1)} \\ F_{(2)} \end{bmatrix} = U_1F_{(1)} + U_2F_{(2)} \tag{4-42}$$

此式称为因子分解式,其中 $U_1F_{(1)}$这一部分是由 p 个主成分能解释的部分,而 $U_2F_{(2)}$就是残余部分,因为

$$R = \frac{1}{N-1}XX^T = \frac{1}{N-1}[U_1F_{(1)} + U_2F_{(2)}][U_1F_{(1)} + U_2F_{(2)}]^T$$

$$= \frac{1}{N-1}[U_1F_{(1)}F_{(1)}^TU_1^T + U_1F_{(1)}F_{(2)}^TU_2^T + U_2F_{(2)}F_{(1)}^TU_1^T + U_2F_{(2)}F_{(2)}^TU_2^T]$$

因此,$R-\frac{1}{N-1}U_1F_{(1)}F_{(1)}^TU_1^T$ 就是残余的协方差矩阵,如果残余协方差矩阵很接近于对角阵,表明这几个主分量选得比较合适。

若记残余部分为 $\varepsilon=[\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_m]^T$,且将 X 写成

$$X = U_1F_{(1)} + \varepsilon \tag{4-43}$$

则(4-43)可以写成如下的模型

$$x_1 = u_{11}F_1 + u_{12}F_2 + \cdots u_{1p}F_p + \varepsilon_1$$

$$x_2 = u_{21}F_1 + u_{22}F_2 + \cdots u_{2p}F_p + \varepsilon_2$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_m = u_{m1}F_1 + u_{m2}F_2 + \cdots u_{mp}F_p + \varepsilon_m$$

上述式子中各因子 $F_1,F_2,\cdots,F_p$ 的方差不一定是 1,当我们要求它们的方差为 1 时,就得到 R 型因子模型如下

$$x_1 = a_{11}F_1 + a_{12}F_2 + \cdots a_{1p}F_p + a_1\varepsilon_1$$

$$x_2 = a_{21}F_1 + a_{22}F_2 + \cdots a_{2p}F_p + a_2\varepsilon_2$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_m = a_{m1}F_1 + a_{m2}F_2 + \cdots a_{mp}F_p + a_m\varepsilon_m \tag{4-44}$$

模型中每个变量都包含有的因子 $F_1,F_2,\cdots,F_p$ 叫做公因子,是相互独立的,可以把它们理解为高维空间中相互垂直的 p 个坐标轴。$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_m$ 是各变量特有的因子,叫做特殊因子。各特殊因子间以及特殊因子与公因子间都是相互独立的。$a_{jj'}$ 是第 j 个变量在第 j' 个公因子上的负荷,叫做因子载荷,由因子载荷组成的矩阵 A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix}$$

叫做因子载荷矩阵。a_j 是 x_j 在特殊因子 ε_j 的负荷,叫特殊因子载荷。

由于我们感兴趣的是主成分能表示的部分,实际应用时常常略去特殊因子部分,于是,因子分析的简化模型是

$$\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1p}F_p \\ x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2p}F_p \\ &\cdots\cdots \\ x_m &= a_{m1}F_1 + a_{m2}F_2 + \cdots + a_{mp}F_p \end{aligned} \tag{4-45}$$

用矩阵形式表示为

$$X = AF \tag{4-46}$$

其中 $X = (x_1, x_2, \cdots, x_p)'$, $F = (F_1, \cdots, F_m)'$。

2. 因子模型中各个量的统计意义与因子模型的几何意义

弄清模型中各个量的统计意义对因子分析结果的解释十分重要。设模型中各因子和特殊因子都是标准化的(均值为零,方差为 1)变量,则各个量具有如下统计意义。

1)因子载荷 $a_{jj'}$ 的统计意义

由式(4-44)有 $x_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jp}F_p + a_j\varepsilon_j$

则

$$x_jF_{j'} = a_{j1}F_1F_{j'} + a_{j2}F_2F_{j'} + \cdots + a_{jp}F_pF_{j'} + a_j\varepsilon_jF_{j'} \tag{4-47}$$

对式(4-47)取数学期望,因各公因子已标准化且相互独立,可得

$$\begin{aligned} E(x_jF_{j'}) &= a_{j1}E(F_1F_{j'}) + a_{j2}E(F_2F_{j'}) + \\ &\cdots + a_{jj'}E(F_jF_{j'}) + \cdots + a_{im}E(F_mF_j) + a_iE(\varepsilon_iF_j) \end{aligned} \tag{4-48}$$

$$r(x_jF_{j'}) = a_{jj'}$$

其中,$r(x_jF_{j'})$表示 x_j 与 $F_{j'}$ 的相关系数。上式说明变量 x_j 在公因子 $F_{j'}$ 上的因子载荷 $a_{jj'}$ 就是 x_j 与 $F_{j'}$ 的相关系数,表征了 x_j 与公因子 $F_{j'}$ 的关联程度。

2)变量共同和 h_j^2 及其统计意义

因子载荷矩阵 A 中各行元素的平方和叫做变量的共同和,第 j 个变量 x_j 的变量共同和是

$$h_j^2 = \sum_{j'=1}^{m} a_{jj'}^2 \qquad (j = 1,2,\cdots,m) \tag{4-49}$$

为了说明其统计意义,现对因子模型中的变量 x_j 计算其方差(用 $D(x_j)$表示)

$$\begin{aligned} D(x_j) &= a_{j1}^2D(F)_1 + a_{j2}^2D(F_2) + \cdots + a_{jp}^2D(F_p) + a_j^2D(\varepsilon_j) \\ &= a_{j1}^2 + a_{j2}^2 + \cdots + a_{jp}^2 + a_j^2 = h_j^2 + a_j^2 \end{aligned} \tag{4-50}$$

但 $D(x_j) = 1$,故有

$$D(x_j) = h_j^2 + a_j^2 = 1 \tag{4-51}$$

即变量 x_j 的方差由两部分组成,第一部分为共同和 h_j^2,它是全部公因子 x_j 的总方差所作的贡献,第二部分是特殊因子产生的方差,称特殊因子方差,仅与 x_j 本身的变化有关,它是使变量 x_j 的方差为 1 的补充值。这样就可以从共同和接近于 1 的程度来说明公因子是否能代表原始

变量的变化。

3)公因子 F_j 的方差贡献 S_j 的统计意义

因子载荷矩阵中各列元素的平方和记为

$$S_j = \sum_{j'=1}^{m} a_{j'j}^2 \qquad (j = 1,2,\cdots,m) \tag{4-52}$$

它是同一公因子 F_j 对诸变量提供的方差贡献之总和,是衡量公因子 F_j 相对重要性的指标。

4)因子载荷矩阵中不同两行的对应元素乘积之和的统计意义

由因子模型有 $x_j x_{j'} = \left(\sum_{k=1}^{m} a_{jk}F_k + a_j\varepsilon_j\right)\left(\sum_{h=1}^{m} a_{j'h}F_h + a_{j'}\varepsilon_{j'}\right)$

当 $j \neq j'$ 时,两边取数学期望后得

$$r_{ij} = \sum_{k=1}^{m} a_{ik}a_{jk} \qquad (i,j = 1,2,\cdots,p \text{ 且 } i \neq j) \tag{4-53}$$

即 x_j 与 $x_{j'}$ 的相关系数($j \neq j'$)就等于 A 的第 j 行与第 j' 行对应元素乘积之和。

5) 因子模型的几何意义

仅对简化因子模型讨论,我们把互不相关的各自方差为 1 的 p 个公因子想象成 p 个互相垂直的单位矢量,以它们为坐标轴就构成了 p 维空间的一个直角坐标系,我们称这些坐标轴为因子轴。这样每个变量 $x_j (j = 1,2,\cdots,m)$ 就可以用 p 维空间的一个向量 x_j 来表示,它在公因子上的载荷就是该向量在这些因子轴上的投影,即

$$a_{jk} = | x_j | \cos(x_j, F_k) \qquad (k = 1,2,\cdots p)$$

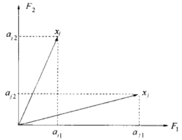


图 4-3 简化因子模型的几何意义

图 4-3 给出了 $p = 2$ 时的示例,x_j 在 F_1 上的载荷为 a_{j1},在 F_2 上载荷为 a_{j2},$x_{j'}$ 在 F_1 上的载荷为 $a_{j'1}$,在 F_2 上载荷为 $a_{j'2}$。图中 x_j 在 F_1 上的载荷较大,也就是 x_j 与 F_1 的相关较密切,此时向量 x_j 与 F_1 的夹角很小,方向趋于一致。

以上所讲的因子模型和因子载荷的几何解释可以给我们一个直观的几何概念或图形,为今后的研究带来方便。

三、Q 型因子分析

前面介绍的 R 型因子分析是研究变量间的相互关系,Q 型因子分析则是研究样品间的相互关系。在 R 型因子分析中用相关系数作为变量间的相似性度量,在 Q 型因子分析中也需要选择样品间合适的相似性度量。这里选择 Imbrie 提出的比例相似系数,从比例相似系数矩阵出发建立了 Imbrie 的 Q 型因子分析。另外也可选择各种距离系数,从距离系数矩阵出发建立 Q 型因子分析。

设 n 个样品 m 个变量的原始数据矩阵为

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = [x_1, x_2, \cdots, x_n]^T \tag{4-54}$$

Q 型因子模型是将 n 个样品用 p 个主因子($F_1, F_2, \cdots, F_p$)表示，其简化的模型是：

$$x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1p}F_p$$
$$x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2p}F_p$$
$$\cdots\cdots$$
$$x_n = a_{n1}F_1 + a_{n2}F_2 + \cdots + a_{np}F_p \tag{4-55}$$

Q 型因子分析的进行步骤完全类似于 R 型因子分析，只不过是从相似矩阵出发而已。

第 i 个样品与第 j 个样品的相似系数为

$$\cos\theta_{ii'} = \frac{\sum_{k=1}^{p} x_{ik}x_{i'k}}{\sqrt{\sum_{m=1}^{p} x_{ik}^2 \sum_{m=1}^{p} x_{i'k}^2}} \tag{4-56}$$

进一步可得相似系数矩阵$[\cos\theta]_{n\times n}$，求出它的前 p 个特征值$\lambda_1, \lambda_2, \cdots, \lambda_p$ 及对应的特征向量 $u_1, u_2, \cdots, u_p$ 则得主因子解 $a_{ij} = u_{ij}\sqrt{\lambda_j}$，$(i = 1,2,\cdots,n; j = 1,2,\cdots,m)$

1.方差最大正交旋转

前面导出了主因子解，无论对 R 型或 Q 型因子分析，我们的目的不仅仅是要找出几个公因子，更重要的是想知道每个公因子代表的地质意义。主因子解只是众多因子解中的一个，或许还有其它因子解更具有易于解释的地质意义。为了找到这样的因子解，我们可以对主因子载荷矩阵进行旋转，使其结构简化，即使每个因子载荷的平方按列方向 0 和 1 两极分化，使第 i 个主因子的代表性样品在 F_i 因子轴上的载荷趋于 1，而在其它因子轴上的载荷趋于 0，这样便于对每个主因子进行地质解释。例如图 4-4a 表示 8 个假想的样品在两个因子轴Ⅰ与Ⅱ上的原始位置，图中看到，它们在因子轴Ⅰ与Ⅱ上的载荷差别不大，因而不便利用因子Ⅰ与Ⅱ对各样品进行解释。但如果将轴Ⅰ与Ⅱ正交旋转一个 θ 角(如图 4-4b)到因子轴Ⅰ'与Ⅱ'的位置，则情况将大大改善，8 个样品的因子载荷开始分化，样品 1、2、3、4 在因子轴Ⅰ'上的载荷较大，而在Ⅱ'上载荷较小，样品 5、6、7、8 在Ⅰ'上的载荷较小而在Ⅱ'上载荷较大。这样，因子Ⅰ'主要反映的是样品 1、2、3、4，因子Ⅱ'主要反映的是样品 5、6、7、8。

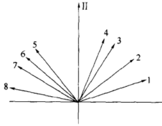


图 4-4a　样品在两个因子轴Ⅰ与Ⅱ上的原始位置

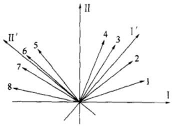


图 4-4b　两个因子轴旋转后的位置

上述旋转变换为正交旋转变换，代数上即对因子载荷阵 A 作正交变换。这里采用的变换称为“方差最大正交旋转变换”，即使因子载荷矩阵 A 中各因子载荷的方差达到最大作为因子载荷矩阵简化的标准。

下边用 Q 型因子模型来说明实现旋转变换的原理和方法，设 Q 型因子载荷矩阵 A 经正交旋转 θ 角后为 B

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix} \qquad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{bmatrix}$$

为避免负值，我们采用因子载荷的平方值，于是第 j 个 $j = 1,2,\cdots,p$ 因子载荷值平方的方差为

$$V_j = \frac{1}{n}\Big[\sum_{i=1}^{n}(b_{ij}^2)^2 - \frac{1}{n}\Big(\sum_{i=1}^{n}b_{ij}^2\Big)^2\Big] \tag{4-57}$$

如果 V_j 为最大，则此时它的组分或趋于 1 或趋于 0，使因子具有最大的简化性。对整个因子载荷矩阵的简化，则可由个因子载荷值平方的方差之和作为衡量的标准。即

$$V' = \sum_{j=1}^{m} V_j = \frac{1}{N}\sum_{j=1}^{m}\Big\{\sum_{i=1}^{N}(b_{ij}^2)^2 - \frac{1}{N}\Big(\sum_{i=1}^{N}b_{ij}^2\Big)^2\Big\} \tag{4-58}$$

考虑到各个样品共同和之间的差异所造成的不平衡，作旋转变换后的 B 矩阵进行规格化处理，即以 b_{ij}^2/h_i^2 代替上式中的 b_{ij}^2，这样就得到

$$V = \frac{1}{n}\sum_{j=1}^{m}\Big\{\sum_{i=1}^{n}(b_{ij}^2/h_i^2)^2 - \frac{1}{n}\Big(\sum_{i=1}^{n}b_{ij}^2/h_i^2\Big)^2\Big\} \tag{4-59}$$

实际上，我们要求经旋转后的 b_{ij} 要使 V 达到极大。

对 A 进行正交旋转，相当于对所有因子面 $F_q - F_l(q = 1,2,\cdots p-1)$；$(l = q+1, q+2,\cdots, p)$ 正交旋转一个角度 θ_{ql}，每次的转角 θ 必须满足使上式中的 V 达到最大，这相当于求一正交矩阵 T，使 $B = AT$ 满足 V 为最大这一条件。设对因子面 $F_q - F_l$ 旋转一角度 θ_{ql} 时的正交矩阵为

$$T_{ql} = \begin{bmatrix} 1 & q & & l & \\ & \vdots & & \vdots & \\ \cdots & \cos\theta_{ql} & \cdots & \sin\theta_{ql} & \cdots \\ & \vdots & & \vdots & \\ \cdots & \sin\theta_{ql} & \cdots & \cos\theta_{ql} & \cdots \\ & \vdots & & \vdots & \\ & & & 1 & \end{bmatrix} \begin{matrix} \\ \\ q \\ \\ l \\ \\ \\ \end{matrix} \tag{4-60}$$

T_{ql} 是 p 阶矩阵，对角线上除 q 行 q 列和 l 行 l 列的元素为 $\cos\theta_{ql}$ 外，其余均为 1，非对角元素除标明的元素以外，其余均为零。

将 A 作正交旋转 θ_{ql} 后得到 $B = AT_{ql} = [b_{ij}]$，B 中的元素 $b_{ij} i = 1,2,\cdots,n j = 1,2,\cdots,m$ 为

$$\begin{aligned} b_{iq} &= a_{iq}\cos\theta_{ql} + a_{il}\sin\theta_{ql} \\ b_{ij} &= -a_{iq}\sin\theta_{ql} + a_{il}\cos\theta_{ql} \qquad (j \neq q, l) \\ b_{ij} &= a_{ij} \end{aligned} \tag{4-61}$$

如果有 p 个主因子，必须将两两配对旋转，共旋转 $p(p-1)/2$ 次，全部旋转完毕算一个循环，此时得因子载荷阵为

$$B_{(1)} = AT_{12}\cdots T_{1p}\cdots T_{(p-1)p} = A\prod_{q=1}^{p-1}\prod_{l=q+1}^{p}T_{ql} = AC_1 \tag{4-62}$$

其中，$C_1 = \prod_{q=1}^{p-1}\prod_{l=q+1}^{p}T_{ql}$ 也是一正交矩阵，$B_{(1)}$ 是对 A 旋行正交变换 C_1 后得出的，对于 $B_{(1)}$

可由(4-59)式算出其 $V_{(1)}$

在第一个循环基础上,从 $B_{(1)}$ 出发在进行第二个旋转循环,然后得 $B_{(2)}$,即

$$B_{(2)} = B_{(1)} \prod_{q=1}^{p-1} \prod_{l=q+1}^{p} T_{ql} = B_{(1)} C_{(2)} = AC_1 C_2 \tag{4-63}$$

从 $B_{(2)}$ 可算出它的 $V_{(2)}$。如此继续下去,不断重复这样的旋转循环,就可以得到 V 值的一个非降序列

$$V_{(1)} \leqslant V_{(2)} \leqslant V_{(3)} \leqslant \cdots \tag{4-64}$$

由于因子载荷的绝对值不大于 1,故该序列是有上界的,因而必收敛于某一极限 $\widetilde{V}$,$\widetilde{V}$ 即是 V 的最大值。也就是说,只要循环次数 k 充分大就会有

$$|V_{(k)} - \widetilde{V}| < \varepsilon \tag{4-65}$$

式中,ε 是给定的精度。

实际计算中,只要第 k 次与第 $k-1$ 次两次循环所得的 $V_{(k)}$ 与 $V_{(k-1)}$ 满足

$$|V_{(k)} - V_{(k-1)}| < \varepsilon \tag{4-66}$$

时即可停止,将最后得到的矩阵 $B_{(k)}$ 作规格化还原,即得旋转后的因子载荷阵。

最后一个问题是如何确定任一次变换 T_{ql} 中的 θ_{ql}(以下简记为 θ),以使 V 经 T_{ql} 变换后达到最大,为此,可将(4-62)式代入(4-60)式,然后再对 θ 求一阶导数令其为 0 即可解得

$$\text{tg}4\theta = \frac{D - 2AB/N}{C - (A^2 - B^2)/N} = \frac{x_n}{x_0}\left(-\frac{\pi}{2} < 4\theta < \frac{\pi}{2}\right) \tag{4-67}$$

其中 $A = \sum_{i=1}^{n} \mu_i$, $B = \sum_{i=1}^{n} \gamma_i$, $C = \sum_{i=1}^{n} (\mu_i^2 - \gamma_i^2)$, $D = 2\sum_{i=1}^{N} \mu_i \gamma_i$, $\gamma_i = 2(a_{iq}/h_i)(a_{il}/h_i)$, $\mu_i = (a_{iq}/h_i)^2 - (a_{il}/h_i)^2$

因为在(4-61)式中实际用到的是 $\sin\theta$ 和 $\cos\theta$,编程中常采用以下式子来求 $\sin\theta$ 和 $\cos\theta$ 的值

$$\cos4\theta = x_0 / \sqrt{x_n^2 + x_0^2} \qquad \cos2\theta = \sqrt{(1 + \cos4\theta)/2}$$

$$\cos\theta = \sqrt{(1 + \cos2\theta)/2} \qquad \sin\theta = sign(x_n)\sqrt{1 - \cos^2\theta} \tag{4-68}$$

至此已完成了从主因子载荷矩阵出发,经方差最大正交旋转,求旋转后的因子载荷矩阵的全过程。

2. 因子得分的计算

上述的 Q 因子模型是将样品表示为公因子的线形组合,这些公因子往往与一定的地质成因有关。反过来,也可以将公因子表示为样品的线形组合,即用

$$F_j = \beta_{1j}x_1 + \beta_{2j}x_2 + \cdots + \beta_{nj}x_n (j = 1,2,\cdots,p) \tag{4-69}$$

来计量各样品的因子得分。对 Q 型分析来说,因子得分是因子中各个变量的取值,可用以研究因子的成分。下边以 Q 型因子分析为例讨论因子得分的计算,R 型因子分析的情况可以类似得出。

如果公因子已提取全部信息的 90% 以上,则可以采用简化的因子模型

$$X = AF \tag{4-70}$$

由于其中 A 是 $n \times p$ 阶矩阵,故先用 A^T 左乘上式两端,然后再用 A^TA 的逆阵左乘两边,即

$$A^TX = A^TAF$$

$$F = (A^T A)^{-1} A^T X \qquad (4-71)$$

这就是因子得分的直接计算公式。

3.因子分析计算步骤

今以 Q 型因子分析为例说明因子分析计算步骤，R 型因子分析与此类似，但应从相关系数矩阵出发。

(1)收集整理好原始数据矩阵。

(2)对原始数据作数据标准化变换。

(3)计算 p 个样品间的相似系数，形成相似系数矩阵 R。

(4)求出 R 的特征值并按大小排列为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$，及相应于 λ_i 的单位特征向量 u_i $(i=1,2,\cdots,p)$。

(5)确定公因子个数 p(按前 p 个特征值之和占特征值总和的百分数大于90%来确定个数 p)。

(6)求出主因子载荷矩阵 $A=[a_{ij}]$，$a_{ij}=u_{ij}\sqrt{\lambda_i}$。

(7)计算共同和 h_i^2、h_i^2 是否接近于1，说明公因子是否已提取了变量 x_i 的大部分变化信息。

(8)将因子载荷矩阵 A 作方差最大正交旋转，求出旋转后的因子载荷矩阵，仍记为 A。

(9)计算因子得分。

四、因子分析的应用实例

下边以 R 型因子分析为例，介绍因子分析的应用。

【例 4-2】 在冀北坳陷东部某烃源岩层采得烃源岩样品4个，每个样品测得三个地化指标，共12个原始数据，将原始数据的各指标变量标准化，得到标准化数据阵 X 为：

$$X = \begin{bmatrix} -1.26 & -0.96 & -0.78 \\ -0.64 & -0.96 & -1.14 \\ 0.640 & 0.580 & 1.320 \\ 1.260 & 1.340 & 0.620 \end{bmatrix}$$

要求对上述标准化数据进行 R 型因子分析，考查各指标变量间关系，计算各主因子的因子计量和因子贡献。

解：(1)计算三个变量的相关系数矩阵

$$R = \frac{1}{4}X^T X = \begin{bmatrix} 1.00 & 0.79 & 0.83 \\ 0.79 & 1.00 & 0.86 \\ 0.83 & 0.86 & 1.00 \end{bmatrix}$$

(2)计算主因子解

首先求得 R 的特征值为 $2.77>0.20>0.03$，方差贡献的累计百分数分别为0.9233,0.99,1.00。

(3)确定公因子数 $m'=2$，计算相应的因子计量

R 前两个特征值所对应的特征向量为如下矩阵的两个行向量

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0.58 & 0.59 & 0.56 \\ -0.47 & -0.33 & 0.83 \end{bmatrix}$$

得到主因子荷载矩阵为

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}\, u_1 \\ \sqrt{\lambda_2}\, u_2 \end{bmatrix} = \begin{bmatrix} 0.96 & 0.98 & 0.93 \\ -0.21 & -0.15 & 0.37 \end{bmatrix}$$

(4)对各主因子贡献的分析

由主因子荷载得到方差贡献分析见表 4－4。由表中的结果看到，各变量的公因子方差都比较大，因此，用两个公因子解释三个变量能取得较好效果，各公因子的方差贡献等于相应的特征值。第一公因子的方差贡献占总方差的 92.3%，但各变量在该公因子上的载荷都很大，在分别为 0.96、0.98 和 0.93，看不出各个变量的亲疏关系。故需要对它进行旋转，求出它的正交因子解，然后进行分析，下边介绍有关的分析内容。

表 4－4　主因子荷载方差贡献分析

因子变量	x_1	x_2	x_3	方差贡献平方和
f_1	0.96	0.98	0.93	2.77
f_2	－0.21	－0.15	0.37	0.20
公因子方差平方和	0.97	0.98	1.0	

(5)计算方差极大正交因子解

要计算方差极大正交因子解，即寻找适当的正交变换矩阵 P，使因子荷载矩阵通过正交矩阵 P 旋转后，得到极少数元素的绝对值接近 1，大多数元素的绝对值接近 0 的简单矩阵结构 B，而且 A 和 B 具有如下关系。其中正交变换矩阵 T 中的旋转角 θ，根据方差极大准则(文献 14)计算。

$$B = PA = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix}$$

$$= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 0.96 & 0.98 & 0.93 \\ -0.21 & -0.15 & 0.37 \end{bmatrix}$$

其中，$\mathrm{tg}\, 4\theta = \dfrac{mD - 2AB}{mC - (A^2 - B^2)}$，$D = 2\sum_{j=1}^{3}[(x_{1j}^2 - x_{2j}^2)(2x_{1j}x_{2j})]$，$A = \sum_{j=1}^{3}(x_{1j}^2 - x_{2j}^2)$，$B = 2\sum_{j=1}^{3} x_{1j}x_{2j}$，$C = \sum_{j=1}^{3}[(x_{1j}^2 - x_{2j}^2)^2 - 2(x_{1j}x_{2j})^2]$

由上述关系计算得到旋转角 $\theta = -69°$，相应的正交矩阵为

$$P = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 0.77 & -0.64 \\ 0.64 & 0.77 \end{bmatrix}$$

$$B = PA = \begin{bmatrix} 0.87 & 0.85 & 0.48 \\ 0.45 & 0.51 & 0.88 \end{bmatrix}$$

(6)计算旋转后的因子计量 F

$$F = XA^T\Lambda_p P^T = \begin{bmatrix} -1.26 & -0.96 & -0.78 \\ -0.64 & -0.96 & -1.14 \\ 0.64 & 0.58 & 1.32 \\ 1.26 & 1.34 & 0.62 \end{bmatrix} \begin{bmatrix} 0.96 & -0.21 \\ 0.98 & -0.15 \\ 0.93 & 0.37 \end{bmatrix} \begin{bmatrix} -2.88 & 0.12 \\ -2.62 & -0.14 \\ 2.41 & 0.27 \\ 3.10 & -0.24 \end{bmatrix} \begin{bmatrix} 0.28 & 0.23 \\ -3.2 & 3.85 \end{bmatrix}$$

$$=\begin{bmatrix} -1.19 & 0.2 \\ -0.29 & -1.24 \\ -0.19 & 1.59 \\ 1.64 & -0.21 \end{bmatrix}$$

(7)正交因子解的地质解释

从旋转后的因子荷载矩阵 $B=\begin{bmatrix} 0.87 & 0.85 & 0.48 \\ 0.45 & 0.51 & 0.88 \end{bmatrix}$，对于变量 x_1 和 x_2 在第一个主成分上的因子贡献较大，而变量 x_3 在第二个主成分上的因子贡献较大，说明烃源岩的这三个地化指标反映了两个不同的烃成熟演化阶段。

另外，从因子计量结果看到，第一和第四块样本在第一个主成分上的得分较大，反映第一个主成分代表的烃成熟时期形成的层段上采取的，而第二和第三块样本是在第二个主成分代表的烃成熟时期形成的层段上采取的。若样本较多且有代表性，利用这些样本的分析结果，可以指导寻找不同时期形成的油气储层。

第三节　聚类分析及其应用

聚类分析又称群分析，是一种数学分类方法。它起源于分类学，是研究分类问题的一种主要方法。自然界中存在大量的分类问题，如生物的分类、经济发展水平的分类和教育水平的分类等。在地质学中也存在多种分类问题，如古生物种族分类、油气圈闭的分类和有机质类型的分类等。以往解决分类问题主要靠专业知识和经验，但随着科学技术的发展，仅靠专业知识和经验已不能适用于精细分类和定量研究的目的。

聚类分析的基本思想是：首先认为所研究的对象存在不同的相似性，根据样品各个指标的观测值，找出一些能够度量样品之间相似程度的统计量，以此为依据，把相似性程度大的样品聚为一类，即将关系密切的样品聚合到一个小的分类单元，关系疏远的又聚到一个大的分类单元，直到把所有样品都聚合完毕，形成一个由小到大的分类系统，最后把分类系统直观地用图形表示出来，形成所谓的谱系图或称为聚类图。

根据分类对象的不同，聚类分析分为两种。一种是对指标(即变量)进行分类，叫做 R 型聚类分析；一种是对样品进行分析，叫做 Q 型聚类分析。

从聚类所用的方法或聚类选用的统计量看，聚类分析可分为系统聚类法、有序聚类法、动态聚类法和模糊聚类法等，本节仅介绍系统聚类法。

在实际工作中，由于样品多个指标的观测值代表同一样品不同属性的量，这些观测值可能在量纲和量级上有较大差别，因此直接用原始数据进行计算，可能会突出某些量级大的指标在分类中的作用，压低甚至淹没某些小量级指标的作用。为此，在聚类计算前需对原始数据进行变换。下面从讨论数据变换方法开始，叙述聚类分析的方法原理、计算步骤和应用实例。

一、原始数据的变换方法

数据变换具有如下目的：(1)统一地质变量的量纲；(2)使地质变量尽可能呈正态分布；(3)使地质变量间的非线性关系转化为线性关系；(4) 用一组新的数量更少的相互独立的变量代替原来一组有相关关系的地质变量。常用的数据变换方法，根据目的不同又有统一量纲的和使数据转换成正态分布的等多种方法，这里仅介绍几种统一量纲的变换方法。

1.标准化变换

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}(i = 1,2,\cdots,n;j = 1,2,\cdots,m) \tag{4-72}$$

式中,$i=1,2\cdots,n$ 为样品数;$j=1,2,\cdots,m$ 为变量数;x_{ij}表示第j个变量在第i个样品中的原始观测值;x'_{ij}为变换后第j个变量在第i个样品中的观测值;$\bar{x}_j$ 是第j个变量的算术平均;s_j 是第 j 个变量的标准差 $s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1}}$。

2.极差变换

$$x'_{ij} = \frac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}}(i = 1,2,\cdots,n;j = 1,2,\cdots,m) \tag{4-73}$$

式中,$x_{j\max}$是第 j 个变量的最大值。$x_{j\min}$是第 j 个变量的最小值。

3.均匀化变换(又称平均数计量变换)

$$x'_{ij} = \frac{x_{ij}}{\bar{x}_j}(i = 1,2,\cdots,n;j = 1,2,\cdots,m) \tag{4-74}$$

为了书写方便,后边我们仍用 x_{ij}表示标准化变量。

二、相似性统计量的定义和计算

聚类分析是以样品指标的相似性程度进行的,用何种相似性度量作为相似性统计量,关系到聚类结果的正确性和实际应用的方便程度。下面给出常用的相似性统计量的定义和计算方法。

1.常用的距离系数

如果把 n 个样品看成m 维空间中的n 个点,则两个样品间相似程度可用 m 维空间中两点的距离来度量,令 $d_{ii'}$ 表示样品S_i 和$S_{i'}$的距离,常用的距离有明氏距离、马氏距离和兰氏距离。

1)明氏(Minkowski)距离

$$d_{ii'}(q) = \left(\sum_{j=1}^{m} | x_{ij} - x_{i'j} |^q\right)^{1/q} \tag{4-75}$$

当 $q=1$ 时, $d_{ii'}(1) = \left(\sum_{j=1}^{m} | x_{ij} - x_{i'j} |\right)$ 为绝对距离;

当 $q=2$ 时, $d_{ii'}(2) = \left(\sum_{j=1}^{m} (x_{ij} - x_{i'j})^2\right)^{1/2}$ 为欧氏距离;

当 $q=\infty$ 时: $d_{ii'}(\infty) = \left(\max_{1\le j\le m} | x_{ij} - x_{i'j} |\right)$ 为切比雪夫距离。

2)马氏(Mahalanobis)距离

马氏距离是印度统计学家马哈拉那比斯于1936年引入的,故称为马氏距离。设$\sum$表示 m 个指标的协方差矩阵

$$\sum = (s_{jj'})_{m\times m} \tag{4-76}$$

其中 $s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$ $(j,j' = 1,2,\cdots,m)$,$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}x_{ij}$,$\bar{x}_{j'} = \frac{1}{n}\sum_{i=1}^{n}x_{ij'}$,如果$\sum^{-1}$存在,则两个样品间的马氏距离为

$$d_{ii'}^2(M) = (X_i - X_{i'})^T \sum^{-1} (X_i - X_{i'}) \tag{4-77}$$

式中，$X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T$，$X_{i'} = (x_{i'1}, x_{i'2}, \cdots, x_{i'm})^T$ 分别表示各个变量在第 i 个样品和第 i' 个样品上的值。

样品 X 到总体 G 的马氏距离定义为：$d^2(X,G) = (X-\bar{x})^T \sum^{-1} (X-\bar{x})$，其中，$\bar{x}$ 为样品的均值，$\sum$ 为 m 个指标协方差矩阵。

3)兰氏(Canberra)距离

由 Lance 和 Williams 最早提出，故称兰氏距离：

$$d_{ii'}(L) = \frac{1}{m}\sum_{j=1}^{m} \frac{|x_{ij} - x_{i'j}|}{x_{ij} + x_{i'j}} (i, i' = 1, 2, \cdots, n) \tag{4-78}$$

需要指出的是，兰氏距离仅适用于一切 $x_{ij} > 0$ 和 $x_{i'j} > 0$ 的情况，这个距离有助于克服指标间的量纲影响，但没有考虑指标间的相关关系。

2. 相关系数

通常所说的相关系数是指变量间的相关系数。设两个指标在某个样品上的值用两个向量 $x_j = (x_{1j}, x_{2j}, \cdots, x_{nj})^T$ 和 $x_{j'} = (x_{1j'}, x_{2j'}, \cdots, x_{nj'})^T$ 表示，则两个指标间的相关系数为

$$r_{jj'} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ij'} - \bar{x}_{j'})^2}} (-1 \leqslant r_{jj'} \leqslant 1) \tag{4-79}$$

其中，$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$，$\bar{x}_{j'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij'}$

样品间的相关系数也可做类似定义，设两个样品的各个指标值分别为 $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T$ 和，$x_{i'} = (x_{i'1}, x_{i'2}, \cdots, x_{i'm})^T$ 则两个样品间相关系数定义为

$$r_{ii'} = \frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2 \sum_{j=1}^{m}(x_{i'j} - \bar{x}_{i'})^2}} (-1 \leqslant r_{ii'} \leqslant 1) \tag{4-80}$$

其中，$\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij}$；$\bar{x}_{i'} = \frac{1}{m}\sum_{j=1}^{m} x_{i'j}$。

三、系统聚类方法

系统聚类分析方法是使分类数由多变少的一种方法。其基本思路是首先认为各对象自成一类；其次，根据两对象间的相似性，将最近的两个对象合并成一类；计算新类与其余各类的相似性，将最近的两类合并成一类。重复上述步骤，直到所有对象合并成一类为止。系统聚类分析中常用的相似程度度量标准有最短距离法、最长距离法和中心距离法，下面就这 3 种方法作一简单介绍。

1. 最短距离法

定义两类 G_p 与 G_q 之间的距离为这两类样品中距离最近的两个样品间的距离，即

$$D_{pq} = \min_{X_i \in G_p, X_{i'} \in G_q} d_{ii'} \tag{4-81}$$

设类 G_p 与 G_q 合并成一个新类记为 G_r，则任一类 G_k 与 G_r 的距离

$$\begin{aligned} D_{kr} &= \min_{X_i \in G_k, X_{i'} \in G_r} d_{ii'} \\ &= \min\{\min_{X_i \in G_k, X_{i'} \in G_p} d_{ii'}, \min_{X_i \in G_k, X_{i'} \in G_q} d_{ii'}\} \\ &= \min\{D_{kp}, D_{kq}\} \end{aligned} \tag{4-82}$$

2.最长距离法

定义类 G_p 与 G_q 之间的距离为两类样品中最远的两个样品间的距离，即

$$D_{ii'} = \max_{X_i \in G_p, X_{i'} \in G_q} d_{ii'} \tag{4-83}$$

设类 G_p 与 G_q 合并成一个新类，记为 G_r，则任一类 G_k 与 G_r 的距离

$$\begin{aligned} D_{kr} &= \max_{X_i \in G_k, X_{i'} \in G_r} d_{ii'} \\ &= \max\{\max_{X_i \in G_k, X_{i'} \in G_p} d_{ii'}, \max_{X_i \in G_k, X_{i'} \in G_q} d_{ii'}\} \\ &= \max\{D_{kp}, D_{kq}\} \end{aligned} \tag{4-84}$$

3.中间距离法

定义类 G_p 与 G_q 之间的距离既不采用两类间的最近距离，也不采用两类间的最远距离，而采用介于两者间的距离，故称中间距离法。

$$D_{kr} = \frac{1}{2}D_{kp} + \frac{1}{2}D_{kq} + \beta D_{pq} \tag{4-85}$$

式(4-85)中，当 $\beta = -\frac{1}{4}$时，D_{kr}就是三角形的中线。

四、谱系图的形成和应用

1.谱系图的形成过程

算出相似性统计量后，就可以它为依据对样品或变量进行聚类，最后形成谱系图。聚合归类时要遵从如下原则：

(1) 若选出的一对样品在已经分好的组中都没出现过，则把它们形成一个独立的新组；

(2) 若选出的一对样品中，有一个出现在已经分好的组中，则把另一个样品也加入到该组中；

(3) 若选出的一对样品，分别出现在已经分好的组中，则把这两个组合并，用谱系图中的线连在一起；

(4) 若选出的一对样品出现在同一组中，则无需再分组。

按上述原则反复进行，直到所有样品聚合分类完毕为止。

2.一步形成法

一步形成法是一种最简单的聚类方法，计算量不大。它由相似性或相似性矩阵出发得到最终的分类结果。这里仅就相似系数作为统计量，介绍这个方法的一般原理：根据相关系数的大小，依次将诸样品归类联结起来，形成一个从小类到大类的分类系统。依次将相关系数最大的两个样品连结成一类，在连结的过程中，要遵循上述4条原则。下面用一个例子说明上述思想的实现过程。

【例4-3】 南海地区取得干酪根样品6个，每个样品测定了6个指标，分别是类脂体、壳

质体、镜质体、惰质体、H/C原子比和O+S/C原子比，原始数据如表4－5所示。现要求对样品进行分类，即进行Q型聚类分析。

表4－5　南海地区干酪根样品指标原始数据

样品/指标	类脂体	壳质体	镜质体	惰质体	H/C	O+S/C
S_1	55	20	25	0	10.0	2.10
S_2	25	10	40	25	15.2	8.13
S_3	15	0	55	30	19.5	10.49
S_4	45	10	40	5	10.1	1.55
S_5	40	20	40	0	10.0	4.67
S_6	45	30	25	0	9.0	2.93

解：（1）数据变换

由于原始数据中各指标量级无多大差别，故不进行数据变换。

（2）样品间的相似性统计量

用样品间的相关系数作相似性统计量，根据式（4－80）算得样品间的相似系数矩阵如下：

$$
\begin{array}{c}
\quad S_1 \quad\; S_2 \quad\; S_3 \quad\; S_4 \quad\; S_5 \quad\; S_6 \\
R = \left[\begin{array}{cccccc}
1.000 & .7524 & .5493 & .9133 & .9519 & .9763 \\
 & 1.000 & .9549 & .8808 & .8617 & .7471 \\
 & & 1.000 & .7654 & .7346 & .5505 \\
 & & & 1.000 & .9790 & .9444 \\
 & & & & 1.000 & .9434 \\
 & & & & & 1.000
\end{array}\right]
\begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array}
\end{array}
$$

以下用一步形成法形成谱系图：

（a）在相似性统计量中找出最大值 $r_{45}=0.9790$，将样品4和5聚合为一类，同时划去R中的第5行和第5列，这里采用的是保留较小标号的第4行和第4列；

（b）在相似性统计量中找出剩余的最大值 $r_{16}=0.9763$，将样品1和6聚合为一类，同时划去R中的第6行和第6列，保留较小标号的第1行和第1列；

（c）在相似性统计量中找出剩余的最大值 $r_{23}=0.9549$，将样品2和3聚合为一类，同时划去R中的第3行和第3列，保留较小标号的第2行和第2列；

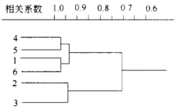


图4－5　样品聚类（Q型聚类）分析后的谱系图

（d）在相似性统计量中找出剩余的最大值 $r_{14}=0.9133$，将样品1和4聚合为一大类，同时划去R中的第4行和第4列，保留较小标号的第1行和第1列；

（e）在相似性统计量中找出剩余的最大值 $r_{12}=0.7524$，将样品1和2聚合为一大类，同时划去R中的第2行和第2列。

聚类过程可总结为表4－6所示的结果，形成的谱系图如图4－3。

表 4-6　聚类过程总结表

联结顺序	联结样品		相似系数	联结顺序	联结样品		相似系数
1	S_4	S_5	0.9790	4	S_1,S_6	S_4,S_5	0.9133
2	S_1	S_6	0.9763	5	S_1,S_6,S_4,S_5	S_2,S_3	0.7524
3	S_2	S_3	0.9549				

3. 多步形成法基本原理和步骤

多步形成法与一步形成法基本步骤相似，唯一的区别是每一步分类后，要把分过类的样品值或指标值合并成为一个新的变量，即在相关系数矩阵 R 中划出最大的元素，将相应的变量合并，赋予新的记号，把合并的变量的数据加权平均作为新变量的数据，再计算合并后新变量与其余各变量的相关系数，建立新的相关阵，再进行下一步分类，重复这一过程直到把所有变量都合并为一类。

与一步形成法相比，多步形成法计算量增加，但分类的精度自然也提高了。

下面的实例展示了利用指标进行多步分类的实施过程，即 R 型聚类分析。

【例 4-4】　从某火山岩区得到火山岩样品 6 个，分析测定 6 种指示喷发和成岩环境的 6 种指示元素的含量列于表 4-7，试用 R 型聚类分析对 6 种指示元素进行分类，分析喷发和成岩过程中指示元素的富集特征及其地质意义。

表 4-7　元素分析的原始数据

指标 / 样品	x_1	x_2	x_3	x_4	x_5	x_6
S_1	2.45	2.35	3.30	2.90	1.90	0.60
S_2	1.90	0.79	3.25	3.50	2.80	1.20
S_3	1.40	0.00	2.70	2.90	2.68	0.47
S_4	2.45	2.17	3.45	3.40	4.00	1.60
S_5	1.95	1.10	3.25	2.00	1.70	0.00
S_6	1.60	0.79	3.10	3.30	2.00	0.65

解：(1)数据标准化，标准化结果列于表 4-8 中。

表 4-8　标准化数据

指标 / 样品	x_1	x_2	x_3	x_4	x_5	x_6
S_1	1.25	1.26	0.53	-0.20	-0.79	-0.30
S_2	-0.15	-0.45	0.32	0.99	0.37	0.87
S_3	-1.42	-1.31	-2.01	-0.20	0.21	-0.55
S_4	1.25	1.06	1.17	0.79	1.91	1.64
S_5	-0.02	-0.11	0.32	-1.99	-1.05	-1.46
S_6	-0.91	-0.45	-0.32	0.61	-0.66	0.20

(2)计算指标间的相关系数

由标准化数据，根据相关系数公式(4－80)计算指标间的相关系数矩阵 R

$$
R^{(1)} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array}
\begin{array}{c}
\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\
\begin{bmatrix}
1.00 & 0.88 & 0.86 & 0.06 & 0.28 & 0.42 \\
 & 1.00 & 0.75 & 0.06 & 0.17 & 0.33 \\
 & & 1.00 & 0.12 & 0.23 & 0.46 \\
 & & & 1.00 & 0.61 & 0.85 \\
 & & & & 1.00 & 0.87 \\
 & & & & & 1.00
\end{bmatrix}
\end{array}
$$

找到最大的相关系数 $r_{x_1x_2}=0.88$，将(x_1,x_2)合并，建立新的相关矩阵

$$
R^{(2)} = \begin{array}{c} \\ x_1x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array}
\begin{array}{c}
\begin{array}{ccccc} x_1x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\
\begin{bmatrix}
1.00 & 0.86 & 0.07 & 0.23 & 0.40 \\
 & 1.00 & 0.12 & 0.23 & 0.46 \\
 & & 1.00 & -0.61 & 0.83 \\
 & & & 1.00 & 0.87 \\
 & & & & 1.00
\end{bmatrix}
\end{array}
$$

得到最大相关系数 $r_{x_5x_6}=0.87$，把(x_5,x_6)合并，建立新的相关矩阵

$$
R^{(3)} = \begin{array}{c} \\ x_1x_2 \\ x_3 \\ x_4 \\ x_5x_6 \end{array}
\begin{array}{c}
\begin{array}{cccc} x_1x_2 & x_3 & x_4 & x_5x_6 \end{array} \\
\begin{bmatrix}
1.00 & 0.86 & 0.07 & 0.33 \\
 & 1.00 & 0.12 & 0.35 \\
 & & 1.00 & 0.76 \\
 & & & 1.00
\end{bmatrix}
\end{array}
$$

得到最大相关系数 $r_{x_1x_2,x_3}=0.86$，把(x_1,x_2)与 x_3 合并记为(x_1,x_2,x_3)，建立新的相关矩阵

$$
R^{(4)} = \begin{array}{c} \\ x_1x_2x_3 \\ x_4 \\ x_5x_6 \end{array}
\begin{array}{c}
\begin{array}{ccc} x_1x_2x_3 & x_4 & x_5x_6 \end{array} \\
\begin{bmatrix}
1.00 & 0.09 & 0.35 \\
 & 1.00 & 0.76 \\
 & & 1.00
\end{bmatrix}
\end{array}
$$

得到最大相关系数 $r_{x_4x_5x_6}=0.76$，把 x_5,x_6 与 x_4 合并记为(x_5,x_6,x_4)，建立新的相关矩阵

$$
R^{(5)} = \begin{array}{c} \\ x_1x_2x_3 \\ x_5x_6x_4 \end{array}
\begin{array}{c}
\begin{array}{cc} x_1x_2x_3 & x_5x_6x_4 \end{array} \\
\begin{bmatrix}
1.00 & 0.27 \\
 & 1.00
\end{bmatrix}
\end{array}
$$

最后将这两类合并成一类，其谱系图如图 4－6 所示。

按上述的相关系数，取 $r=0.75$ 的水平，可将上述指示元素分成两类，(x_1,x_2,x_3) 为一类，代表喷发过程的指示元素，(x_5,x_6,x_4)为另一类，代表火山岩成岩改造产生的指示元素。

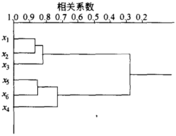


图 4－6 指标聚类(R 型聚类)分析得到的谱系图

第四节　判别分析及其应用

一、判别分析概述

判别分析是判别样品所属类型的一种统计方法。与聚类分析不同，判别分析的前提是判别的几种类型是已知的，而且每一类型都有一批已知样品。判别分析的主要思想是用统计方法将待判的未知样品与已知类型的样品进行类比，以确定待判样品应归属于哪一类。

石油与天然气地质中有大量判别类型的问题，如判别岩石类型、地层年代、古生物种群、钻遇地层的含油气性的好坏、沉积相类型和生油条件的优劣等。因此，在石油地质学中判别分析有着广泛的应用。最简单的情形下，单就一个变量的值就可区分不同的类，见图 4－7(a)。但通常情况下，单就一个变量难于区分不同的类，如图 4－7(b)；只有利用多个变量，才能比较全面地从各个不同的角度刻画各母体间的差异。

实际工作中，由于各个变量间的关系错综复杂，凭简单的逻辑思考进行直观推断，不能充分合理地利用它们所提供的分类依据。下面介绍的判别分析方法就是为解决这个矛盾，用适当的数学方法从多个变量中提出有关判别分类的有效信息，综合成一个像单变量那样好用的判别标准。下面以两组判别为例，说明判别分析的思想。

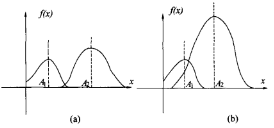


图 4－7　单变量判别分析的示意图

设某区域内存在含矿和不含矿的两组岩体，若把含矿岩体的总体记为 A_1，不含矿岩体的总体记为 A_2，现分别从 A_1 和 A_2 两类总体中个取一批样品，测出样品中能反映含矿特征的一些指标，从所测得的数据看，A_1 和 A_2 的样品数据是有区别的，但也有一部分是相互重叠的。那么，是否可由这各个指标综合出一个新的指标，使得 A_1 和 A_2 在综合指标上有明显的区别？若可以，当取得新的样品时，便可以由新样品的综合指标来判别它属于 A_1 或属于 A_2。事实上，这样的指标对大多数问题是可以找到的。下面只用简单的两组判别问题为例，以图 4－8 的几何图形加以说明。

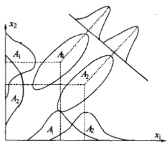


图 4－8　双变量判别分析的示意图

设对 A_1 和 A_2 观测了两个指标 x_1 和 x_2，两个椭圆分布分别是 A_1 组和 A_2 组样品测得的指标的分布，从图 4－8 可见，如果用单一的指标，则 A_1 组和 A_2 组的指标分布有较大的重叠，但如果经过原点作一适当的直线，如式(4－86)所示，将指标分布投影到该直线上，则在该直线上 A_1 和 A_2 两组的指标的重叠部分将小于其它任何方向的直线上的投影。为便于看清，图4－6中用的是该直线的平行线上的投影来说明问题。过原点直线的方程由两个指标 x_1 和 x_2 的线性组合构成，即

$$z = c_1x_1 + c_2x_2 = \sum_{j=1}^{2} c_jx_j \tag{4-86}$$

式中，z 是综合 x_1 和 x_2 两个指标的一个综合指标，称为判别函数，c_1 和 c_2 为判别系数。

由于判别函数是一个线性函数，通常称为线性判别函数。该函数将平面分成两个不相交的区域，即 $c_1x_1 + c_2x_2 \geqslant z_0$ 和 $c_1x_1 + c_2x_2 < z_0$，从而能有效地区分两组样品。

除了线性判别函数外，还有非线性判别函数。实际工作中，可用不同的原则建立判别函数，常用的判别函数有距离判别函数、费歇尔判别函数和贝叶斯判别函数。我们可根据未知样品的指标和分类性质，选用合适的判别函数。下面仅介绍费歇尔判别和贝叶斯判别的方法原理和实现过程。

二、费歇尔准则下的两组判别

1. 费歇尔准则

最为简单的判别分析问题是两组判别问题，如在砂体存在和不存在的样品类属间判别、在含矿和不含矿间的样品类属间判别以及在含油气和不含油气圈闭间的判别等。

设有两个总体分别记为 A_1 和 A_2，它们的样品数分别为 n_1 和 n_2，测得的指标个数为 m，则两类样品的数据可由如下两个数据阵表示

$$A_1 = \begin{bmatrix} x_{11}(A_1) & x_{12}(A_1) & \cdots & x_{1m}(A_1) \\ x_{21}(A_1) & x_{22}(A_1) & \cdots & x_{2m}(A_1) \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_1 1}(A_1) & x_{n_1 2}(A_1) & \cdots & x_{n_1 m}(A_1) \end{bmatrix} \tag{4-87}$$

$$A_1 = \begin{bmatrix} x_{11}(A_2) & x_{12}(A_2) & \cdots & x_{1m}(A_2) \\ x_{21}(A_2) & x_{22}(A_2) & \cdots & x_{2m}(A_2) \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_2 1}(A_2) & x_{n_2 2}(A_2) & \cdots & x_{n_2 m}(A_2) \end{bmatrix} \tag{4-88}$$

设区分这两个样品类属所求得的判别函数为

$$z = c_1x_1 + c_2x_2 + \cdots + c_mx_m = \sum_{j=1}^{m} c_jx_j \tag{4-89}$$

这相当于在 m 维空间中，用 $m-1$ 维超平面将两组样品区分开来。求式(4-89)中的判别系数是求判别函数的关键。

根据上述两组判别函数，可以得到 A_1 和 A_2 两组数据的均值和相应的离散度，即

A_1 组的组内均值

$$\bar{z}(A_1) = \sum_{j=1}^{m} c_j\bar{x}_j(A_1), \bar{x}_j(A_1) = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}(A_1)(j = 1,2,\cdots,m) \tag{4-90a}$$

A_2 组的组内均值

$$\bar{z}(A_2) = \sum_{j=1}^{m} c_j\bar{x}_j(A_2), \bar{x}_j(A_2) = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{ij}(A_2)(j = 1,2,\cdots,m) \tag{4-90b}$$

两组的离散度分别为

$$D(A_1) = \sum_{i=1}^{n_1}[z_i(A_1) - \bar{z}(A_1)]^2 \tag{4-91a}$$

$$D(A_2) = \sum_{i=1}^{n_2}[z_i(A_2) - \bar{z}(A_2)]^2 \tag{4-91b}$$

式中，$z_i(A_1) = \sum_{j=1}^{m} c_j x_{ij}(A_1)(i = 1,2,\cdots,n_1)$；$z_i(A_2) = \sum_{j=1}^{m} c_j x_{ij}(A_2)(i = 1,2,\cdots,n_1)$。

根据判别分析的思想，我们要使两组的重叠部分尽可能小，则要求 z 的值在两组间的差尽可能大，各组内的离散度尽可能小，这就是费歇尔判别准则。

两组的差异可表示为

$$Q = [\bar{z}(A_1) - \bar{z}(A_2)]^2 \tag{4-92}$$

而两组内部的离散度可表示为

$$F = \sum_{i=1}^{n_1}[z_i(A_1) - \bar{z}(A_1)]^2 + \sum_{i=1}^{n_2}[z_i(A_2) - \bar{z}(A_2)]^2 \tag{4-93}$$

按费歇尔准则，应确定系数 $c_1, c_2, \cdots, c_m$ 使 $I = \frac{Q}{F}$ 达到极大。

2. 判别系数的确定

根据极值计算原理，要使 $I = \frac{Q}{F}$ 达到极大，即要求 I 对判别系数 $c_1, c_2, \cdots, c_m$ 的一阶导数为 0，亦即要求 $\frac{\partial I}{\partial c_j} = 0 \quad (j = 1,2,\cdots,m)$，下面具体导出系数的计算关系式

$$\frac{\partial I}{\partial c_j} = \frac{\frac{\partial Q}{\partial c_j}F - Q\frac{\partial F}{\partial c_j}}{F^2} = 0 \tag{4-94}$$

$$\frac{\partial Q}{\partial c_j}F = Q\frac{\partial F}{\partial c_j}, \text{即}\frac{\partial F}{\partial c_j} = \frac{1}{I}\frac{\partial Q}{\partial c_j} \tag{4-95}$$

下面将 I、Q 和 F 的关系式代入式(4－95)的右端和左端项，并将未知数的下标 j 换成 j' 得

$$\begin{aligned}\frac{\partial F}{\partial c_{j'}} &= 2\sum_{g=1}^{2}\sum_{i=1}^{n_g}\left\{\sum_{j=1}^{m} c_j[x_{ij}(A_g) - \bar{x}_j(A_g)]\right\}[x_{ij'}(A_g) - \bar{x}_{j'}(A_g)] \\ &= 2\sum_{j=1}^{m} c_j\left\{\sum_{g=1}^{2}\sum_{i=1}^{n_g}[x_{ij}(A_g) - \bar{x}_j(A_g)][x_{ij'}(A_g) - \bar{x}_{j'}(A_g)]\right\} \\ &= 2\sum_{j=1}^{m} w_{j'j}c_j \quad (j' = 1,2,\cdots,m)\end{aligned} \tag{4-96}$$

式中，$w_{j'j} = \sum_{g=1}^{2}\sum_{i=1}^{n_g}[x_{ij}(A_g) - \bar{x}_j(A_g)][x_{ij'}(A_g) - \bar{x}_{j'}(A_g)]$。

将 A_1 组的均值和 A_2 组的均值代入式(4－92)得

$$\begin{aligned}\frac{\partial Q}{\partial c_{j'}} &= 2\left\{\sum_{j=1}^{m} c_j[\bar{x}_j(A_1) - \bar{x}_j(A_2)]\right\}[\bar{x}_{j'}(A_1) - \bar{x}_{j'}(A_2)] \\ &= 2\left(\sum_{j=1}^{m} c_j d_j\right)d_{j'} \quad (j' = 1,2,\cdots,m)\end{aligned} \tag{4-97}$$

式中，$d_j = [\bar{x}_j(A_1) - \bar{x}_j(A_2)]$。

将式(4－96)、(4－97)代入方程(4－95)得

$$2\sum_{j=1}^{m} w_{j'j}c_j = 2\frac{\sum_{j=1}^{m} c_j d_j}{I}d_{j'} \quad (j' = 1,2,\cdots,m) \tag{4-98}$$

若 $c_1, c_2, \cdots, c_m$ 是这一方程组的解，则对任意常数 $\alpha \neq 0$，$\alpha c_1, \alpha c_2, \cdots, \alpha c_m$ 也是它的一组

解,我们可以取适当的 $\alpha \neq 0$,使 $\frac{\sum_{j=1}^{m} \alpha c_j d_j}{I} = 1$,从而通过求下面的方程组求取判别系数

$$\sum_{j=1}^{m} w_{j'j} c_j = d_{j'} \quad (j' = 1,2,\cdots,m) \tag{4-99}$$

简写成矩阵形式为

$$WC = D \tag{4-100}$$

式中,$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mm} \end{bmatrix}$,系数矩阵 W 称为 m 个变量的类内离差矩阵;

$D = [d_1 \quad d_2 \quad \cdots \quad d_m]^T, C = [c_1 \quad c_2 \quad \cdots \quad c_m]^T$。

求取了判别系数,即得到了式(4-89)的判别函数,将未知类的样品值代入判别函数即可实现对未知类的判别。

3. *判别规则和判别效果检验*

求取判别函数后,可以用公式(4-90)计算出组内平均并确定判别指标,具体可以采用如下两种方法:

(1)由各组样品数 n_1, n_2,取 z_0 为 $\bar{z}_1$ 和 $\bar{z}_2$ 的加权平均,即计算判别函数的总平均

$$z_0 = \frac{n_1 \bar{z}_1 + n_2 \bar{z}_2}{n_1 + n_2} = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} z_i(A_1) + \sum_{i=1}^{n_2} z_i(A_2)\right] \tag{4-101}$$

(2)求取两类均值之间的欧氏距离。

用上面的方法确定的判别函数,当 n_1, n_2 较大时,受样品数影响较大,往往出现偏差。

上面分析的判别方法只考虑了两组样品确有差异的情形。若两组样品没有差异,那么上述讨论就失去了意义。在实际工作中,还需要判别研究的两组是否真有差异。显然要进行样品差异的显著性检验。我们知道,向量 $D = [d_1 \quad d_2 \quad \cdots \quad d_m]^T$ 反映两类之间的差异。为了度量这个差异的大小,这里采用两组均值之间的欧氏距离

$$D^T D = \sum_{j=1}^{m} d_j^2 = \sum_{j=1}^{m} [\bar{x}_j(A_1) - \bar{x}_j(A_2)]^2 \tag{4-102}$$

在式(4-102)中,没有考虑变量间的相关性,而把各变量同等对待,有些情况下与实际数据特性不符。

为消除变量间相关性的影响,需采用马哈拉诺比斯(Mahalanobis)广义距离的概念来度量组间的差异。回顾上一节介绍的马氏距离的定义,首先计算变量间的协方差

$$\begin{aligned} s_{jj'} &= \frac{1}{n-2}\sum_{g=1}^{2}\sum_{i=1}^{n_g}[x_{ij'}(A_g) - \bar{x}_{j'}(A_g)][x_{ij}(A_g) - \bar{x}_j(A_g)] \\ &= \frac{1}{n-2} w_{jj'} \quad (j,j' = 1,2,\cdots,m) \end{aligned} \tag{4-103}$$

式中,$n = n_1 + n_2$。

m 个变量的协方差矩阵定义为

$$S = [s_{jj'}]_{m \times m} = \frac{1}{n-2} W, S^{-1} = (n-2) W^{-1} \tag{4-104}$$

一般地，S 和 S^{-1} 都为正定，两类均值的马氏距离可定义为如下的二次形

$$D_{ma}^2 = D^T S^{-1} D \tag{4-105}$$

式中，D_{ma} 表示马氏距离。

利用马氏距离进行检验的统计量可定义为

$$F = \frac{n-m-1}{m} \times \frac{n_1 n_2}{n(n-2)} D_{ma}^2 \tag{4-106}$$

当两组样品的均值相等时，它服从自由度为 $(m, n-m-1)$ 的 F 分布，用它可检验两组的均值是否存在差异。

4. 实例分析

【例 4-5】现有表 4-9 所示的若干个油层和水层的样品的 4 个指标：岩性系数(x_1)、孔隙度(x_2)、侵入系数(x_3)和含油气饱和度(x_4)，若油层为 A_1 组，水层为 A_2，现要求根据表中的数据建立两组的判别函数并检验其有效性。

表 4-9　油层和水层样品指标

油层	x_1	x_2	x_3	x_4	z
1	0.276	0.18	0.446	0.683	11.60
2	0.378	0.20	0.746	0.673	8.05
3	0.325	0.20	0.800	0.633	7.75
4	0.138	0.21	0.750	0.728	15.09
5	0.290	0.24	0.870	0.649	7.43
6	0.270	0.19	1.730	0.613	8.51
7	0.450	0.23	2.660	0.544	0.07
8	0.302	0.23	1.780	0.590	5.27
9	0.344	0.24	3.400	0.618	4.67
10	0.358	0.21	1.370	0.619	5.98
11	0.076	0.26	0.850	0.733	14.59
12	0.346	0.27	1.320	0.621	3.85
13	0.186	0.30	0.560	0.796	12.85
水层	x_1	x_2	x_3	x_4	z
1	0.62	0.24	6.22	0.544	-4.78
2	0.61	0.25	1.42	0.494	-6.01
3	0.62	0.27	1.46	0.510	-6.47
4	0.56	0.13	1.30	0.372	-4.43
5	0.43	0.22	0.90	0.214	-10.85
6	0.47	0.20	2.90	0.220	-11.20
7	0.56	0.20	3.00	0.221	-13.21
8	0.29	0.25	4.66	0.395	-3.01
9	0.30	0.22	3.18	0.250	-7.17
10	0.35	0.19	17.90	0.230	-10.19
11	0.27	0.25	8.70	0.145	-12.50

解:(1)计算各组指标的平均值和均值差

$\bar{x}_1(A_1) = 0.2876, \bar{x}_2(A_1) = 0.2278, \bar{x}_3(A_1) = 1.3294, \bar{x}_4(A_1) = 0.6538,$

$\bar{x}_1(A_2) = 0.4618, \bar{x}_2(A_2) = 0.2195, \bar{x}_3(A_2) = 4.6945, \bar{x}_4(A_2) = 0.3268,$

$d_1 = -0.1742, d_2 = 0.0083, d_3 = -3.3651, d_4 = 0.327$

(2)计算两组的综合协方差矩阵并形成方程组

由方程(4-103)计算得到的综合协方差矩阵为

$$W = \begin{bmatrix} 0.0147 & -0.0006 & -0.1209 & 0.0030 \\ -0.0006 & 0.0013 & -0.0004 & 0.0013 \\ -0.1209 & -0.0004 & 11.7019 & -0.1153 \\ 0.0030 & 0.0013 & -0.1153 & 0.0115 \end{bmatrix}$$

得到方程组为

$$WC = D$$

式中,$C = [c_1, c_2, c_3, c_4]^T, D = [d_1, d_2, d_3, d_4]^T$。

(3)解上述方程组

$$c_1 = -22.5698, c_2 = -41.3115, c_3 = -0.17016, c_4 = 37.1$$

从而,判别函数为

$$z = -22.5698x_1 - 41.3115x_2 - 0.17016x_3 + 37.1x_4$$

(4)计算各组判别函数的平均值和判别指标

$$\bar{z}(A_1) = \sum_{j=1}^{4} \bar{x}_j(A_1) = 8.1307, \bar{z}(A_2) = \sum_{j=1}^{4} \bar{x}_j(A_2) = -8.1667$$

$$z_0 = \frac{13 \times 8.1307 + 11 \times (-8.1667)}{13 + 11} = 0.661$$

(5)计算马氏距离和统计量 F

$$D_{Ma}^2 = \sum_{j=1}^{4} c_j d_j = 16.30$$

$$F = \frac{13 \times 11}{(13 + 11)(13 + 11 - 2)} \times \frac{13 + 11 - 4 - 1}{4} \times 16.30 = 21$$

在 $\alpha = 0.05$ 的显著性水平下,查附表 4 得 $F_{0.025}(4,19) = 3.56$。

因 $F > F_{0.025}(4,19)$,显然判别函数是高度显著的。

在上例中加入表 4-10 中的样品后,用费歇尔准则进行两组判别,判别过程如下:

表 4-10 油水层样品指标表

样品	x_1	x_2	x_3	x_4
1	0.36	0.19	3.8	0.56
2	0.42	0.14	0.84	0.54
3	0.36	0.29	4.20	0.50
4	0.35	0.17	3.18	0.61
5	0.32	0.30	5.20	0.62
6	0.52	0.27	3.00	0.58
7	0.61	0.18	1.20	0.59

a.新样品的判别函数值为

$$z_1 = 4.16; \quad z_2 = 4.63; \quad z_3 = -2.27$$

$$z_4 = 7.17; \quad z_5 = 2.50; \quad z_6 = -1.88; \quad z_7 = 0.48$$

b.根据新样品的判别函数值与判别指标的关系,确定各新样品的归属。

判别函数大于 $z_0 = 0.661$ 的样品有 1 号、2 号、4 号和 5 号 4 个样品,故这 4 个样品应归入到油层。

判别函数小于 $z_0 = 0.661$ 的样品有 3 号、6 号和 7 号 3 个样品,故这 3 个样品应归入到水层。

三、贝叶斯准则下的多组判别

前面介绍了两组判别方法,但在实际工作中往往会遇到多组判别的问题,如油气储层研究中油气、油气水同层和水层的问题;根据岩石地球化学资料判别多个沉积环境的问题等,下面介绍在贝叶斯判别准则下的多组判别。

1.贝叶斯判别准则

设有 k 个组 $A_1, A_2, \cdots, A_k$,其中第 g 个组取得的样品个数为 $n_g(g = 1,2,\cdots,k)$,每个样品都观测了 m 个指标,用 $x_{ij}(A_g)$表示第 g 组第 i 个样品的第 j 项指标的观测值,即这 k 个组的观测数据可用如下的数据阵表示

$$A_1 = \begin{bmatrix} x_{11}(A_1) & x_{12}(A_1) & \cdots & x_{1m}(A_1) \\ x_{21}(A_1) & x_{22}(A_1) & \cdots & x_{2m}(A_1) \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_1 1}(A_1) & x_{n_1 2}(A_1) & \cdots & x_{n_1 1}(A_1) \end{bmatrix} \qquad (4-107a)$$

$$A_2 = \begin{bmatrix} x_{11}(A_2) & x_{12}(A_2) & \cdots & x_{1m}(A_2) \\ x_{21}(A_2) & x_{22}(A_2) & \cdots & x_{2m}(A_2) \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_2 1}(A_2) & x_{n_2 2}(A_2) & \cdots & x_{n_2 1}(A_2) \end{bmatrix} \qquad (4-107b)$$

……

$$A_k = \begin{bmatrix} x_{11}(A_k) & x_{12}(A_k) & \cdots & x_{1m}(A_k) \\ x_{21}(A_k) & x_{22}(A_k) & \cdots & x_{2m}(A_k) \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_k 1}(A_k) & x_{n_k 2}(A_k) & \cdots & x_{k_k 1}(A_k) \end{bmatrix} \qquad (4-107c)$$

现有一新样品 Y 来自上述几个组中的某一个,其 m 项指标为 $y_1, y_2, \cdots, y_m$,可用 $Y = (y_1, y_2, \cdots, y_m)$表示,判别分析要解决的问题是将 Y 归于上述的哪一组。

在贝叶斯意义下建立判别模型,是要计算 Y 属于各组的概率,然后比较 $P\{1/Y\}, P\{2/Y\}, \cdots, P\{k/Y\}$的大小,最后将 Y 归于概率最大的那一组。

根据贝叶斯公式,样品 Y 属于第 g 组的概率 $P\{g/Y\}$为

$$P\{g/Y\} = \frac{q_g f_g(y_1, y_2, \cdots, y_m)}{\sum_{l=1}^{k} q_l f_l(y_1, y_2, \cdots, y_m)} \quad (g = 1,2,\cdots k) \qquad (4-108)$$

式中，$f_g(y_1,y_2,\cdots,y_m)$是样品 Y 属于第 g 组的概率密度；q_g 为第 g 组的先验概率，在实际工作中常用样本的频率作为它的估计值。

$$q_g = n_g/N \quad (N\text{ 为全部样品数}, N = \sum_{g=1}^{k} n_g) \qquad (4-109)$$

下面就样本是正态分布时的情况讨论 Y 属于第 g 组的概率 $P\{g/Y\}$ 的计算问题。

2.样品为正态分布时的线性判别函数计算

设 k 组样品服从正态分布，其中第 g 个组的均值为 $\bar{X}_g$ 向量，协方差矩阵为 S_g，于是其概率密度为

$$f_g(x_1,x_2,\cdots,x_m) = \frac{|S_g|^{-1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(X_g-\bar{X}_g)^T S_g^{-1}(X_g-\bar{X}_g)\right] \quad (g=1,2,\cdots,k) \qquad (4-110a)$$

式中，X_g 是第 g 组 n_g 个样品上的 m 个指标组成的数据阵 $X_g = (x_1^g, x_2^g, \cdots, x_j^g, \cdots, x_m^g)$，$x_j^g = (x_{1j}^g, x_{2j}^g, \cdots, x_{m_g j}^g)$；$\bar{X}_g = (\bar{X}_1^g, \bar{X}_2^g, \cdots, \bar{X}_j^g, \cdots \bar{X}_m^g)$ 是均值向量，$\bar{X}_j^g = \frac{1}{n_g}\sum_{i=1}^{n_g} x_{ij}(A_g)$；$S_g = \begin{bmatrix} S_{11}^g & S_{12}^g & \cdots & S_{1m}^g \\ S_{21}^g & S_{22}^g & \cdots & S_{2m}^g \\ \vdots & \vdots & & \vdots \\ S_{m1}^g & S_{m2}^g & \cdots & S_{mm}^g \end{bmatrix}$是协方差矩阵，$S_{jj'}^g = \frac{1}{n_g}\sum_{i=1}^{n_g}[x_{ij}(A_g)-\bar{X}_j^g][x_{ij'}(A_g)-\bar{X}_{j'}^g]$。

对于各协方差函数相等的情况，用 S 代替 S_g，式(4-110a)变为

$$f_g(x_1,x_2,\cdots,x_m) = \frac{|S|^{-1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(X_g-\bar{X}_g)^T S^{-1}(X_g-\bar{X}_g)\right]$$

$$(g=1,2,\cdots k) \qquad (4-110b)$$

将式(4-110a)代入式(4-108)得

$$P\{g/Y\} = \frac{q_g\exp\left[-\frac{1}{2}(Y-\bar{X}_g)^T S^{-1}(Y-\bar{X}_g)\right]}{\sum_{l=1}^{k} q_l\exp\left[-\frac{1}{2}(Y-\bar{X}_l)^T S^{-1}(Y-\bar{X}_l)\right]} \qquad (4-111)$$

利用上述公式即可计算样品 Y 属于各组的概率，然而我们关心的是哪个最大，由上式可以看到分子达到最大即可，下面仅对分子项进行分析。不妨对分子取对数

$$\ln\left\{q_g\exp\left[-\frac{1}{2}(Y-\bar{X}_g)^T S^{-1}(Y-\bar{X}_g)\right]\right\}$$

$$= \ln q_g - \frac{1}{2}(Y-\bar{X}_g)^T S^{-1}(Y-\bar{X}_g) \qquad (4-112)$$

$$= \ln q_g - \frac{1}{2}[Y^T S^{-1} Y - Y^T S^{-1}\bar{X}_g - \bar{X}_g^T S^{-1} Y + \bar{X}_g^T S^{-1}\bar{X}_g]$$

将上式中与 g 有关的项记作 $F_g(Y)$，即

$$F_g(Y) = \ln q_g + \frac{1}{2}Y^T S^{-1}\bar{X}_g + \frac{1}{2}\bar{X}_g^T S^{-1} Y - \frac{1}{2}\bar{X}_g^T S^{-1}\bar{X}_g \quad (g=1,2,\cdots,k) \qquad (4-113)$$

从公式(4-113)可以看到，$F_g(Y)$是 $y_1, y_2, \cdots y_m$ 的线性函数，显然 $F_g(Y)$ 极大时，$P\{g/Y\}$也达到极大，样品 Y 就属于那一组，称 $F_g(Y)$为判别函数。

对于多个变量的判别效果检验，请感兴趣的读者参考相关文献。

四、逐步判别分析

在判别分析中,人们希望用尽可能多的变量构造判别函数,使之能包含尽可能多的与判别分组有关的信息。和前面分析的回归分析方法一样,若包含过多的变量,有时反而会使判别效果变坏。其原因是,在求判别函数时需对协方差矩阵 S 求逆,若包含的变量太多,S 求逆的计算精度较差。特别是当所含变量接近线性相关时,S 接近降秩矩阵,求逆结果极不稳定,判别效果会很差。因此也要像逐步回归那样,在可供选择的诸多变量中选择判别能力强的变量,剔除判别能力弱的变量,这就是逐步判别分析的思想。下面还是以式(4-107)中的样品数据为出发点介绍逐步判别分析的计算步骤。

1.离差矩阵和 Wilks 统计量计算

组内离差矩阵 $W = [w_{jj'}]_{m\times m}, w_{jj'} = \sum_{g=1}^{k}\sum_{i=1}^{n_g}[x_{ij}(A_g) - \bar{x}_j(A_g)]\cdot[x_{ij'}(A_g) - \bar{x}_{j'}(A_g)]$

组间离差矩阵 $B = [b_{jj'}]_{m\times m}, b_{jj'} = \sum_{g=1}^{k} n_g[x_{ij}(A_g) - \bar{x}_{oj}]\cdot[x_{ij'}(A_g) - \bar{x}_{oj'}(A_g)]$

总离差矩阵 $T = [t_{jj'}]_{m\times m}, t_{jj'} = w_{jj'} + b_{jj'}$

其中,$\bar{x}_{oj} = \frac{1}{N}\sum_{g=1}^{k}\sum_{i=1}^{n_g} x_{ij}^g, \bar{x}_{oj'} = \frac{1}{N}\sum_{g=1}^{k}\sum_{i=1}^{n_g} x_{ij'}^g (j = 1,2,\cdots m; j' = 1,2,\cdots,m), N = n_1 + n_2 + \cdots + n_k$,则 Wilks 统计量由矩阵 W 和 T 的行列式值之比得到,记为 U

$$U = |W|/|T| \tag{4-114}$$

2.逐步筛选变量

设已计算了 L 步,在判别式中引入了 l 个变量,则在第$(L+1)$步的计算内容如下:

(1)计算全部变量的判别检验量,对已选入的变量 x_j 和未入选的变量 $x_{j'}$ 分别计算

$$\begin{aligned} U_j &= w_{jj}^{(L)}/t_{jj}^{(L)} \\ U_{j'} &= w_{j'j'}^{(L)}/t_{j'j'}^{(L)} \end{aligned} \tag{4-115}$$

(2) 在已入选的变量中,考察是否有判别能力不显著的变量,这时可在 U_j 中找出最大者(即对应判别力最差的变量),若 $U_r = \max\{U_j\}$,作 F 检验,计算

$$F_1 = \frac{w_{rr}^{(L)} - t_{rr}^{(L)}}{t_{rr}^{(L)}} \times \frac{N - m - l + 1}{m - 1} \tag{4-116}$$

如果 $F_1 \leqslant F_{\alpha/2}(m-1, N-m-l-1)$或 $F_1 > F_{1-\alpha/2}(m-1, N-m-l-1)$,则认为在 α 检验水平下,x_r 的判别能力不显著,从而剔除 x_r,转到第 3 步。

若 $F_1 > F_{\alpha/2}(m-1, N-m-l-1)$或 $F_1 \leqslant F_{1-\alpha/2}(m-1, N-m-l-1)$,则认为在 α 检验水平下,x_r 的判别能力显著,从而保留 x_r,然后转而考察未选入变量中的其它变量是否还有判别能力显著的变量未被选入判别函数。为此,对未被选入的变量找出 $U_{j'}$ 中的最小者,它对应判别能力最强的变量。若 $U_p = \min\{U_{j'}\}$,据此计算 F_2 统计量

$$F_2 = \frac{t_{pp}^{(L)} - W_{pp}^{(L)}}{W_{pp}^{(L)}} \times \frac{N - m - l}{m - 1} \tag{4-117}$$

若 $F_2 > F_{\alpha/2}(m-1, N-m-l)$或 $F_2 \leqslant F_{1-\alpha/2}(m-1, N-m-l)$,则认为在 α 检验水平下,

x_p 的判别能力显著,从而将 x_p 引入到判别函数,转到第 3 步。若 $F_2 \leqslant F_{\alpha/2}(m-1, N-m-l)$ 或 $F_2 > F_{1-\alpha/2}(m-1, N-m-l)$,则 x_p 的判别能力不足,没有变量可再引入判别函数了。重复以上计算,直到再没有可被引入的变量为止,筛选变量环节结束。

3.矩阵 W 和 T 的变换公式

在第 $L+1$ 步对矩阵 W 和 T 作消去 r 列的变换公式如下:

$$w_{jj'}^{(L+1)} = \begin{cases} w_{rj'}^{(L)} / w_{rr}^{(L)} & (j = r, j' \neq r) \\ w_{jj'}^{(L)} - w_{j}^{(L)} w_{rj}^{(L)} / w_{rr}^{(L)} & (j \neq r, j' \neq r) \\ 1 / w_{rr}^{(L)} & (j = r, j' = r) \\ - w_{jr}^{(L)} / w_{rr}^{(L)} & (j \neq r, j' = r) \end{cases} \tag{4-118}$$

$$t_{jj'}^{(L+1)} = \begin{cases} t_{rj'}^{(L)} / t_{rr'}^{(L)} & (j = r, j' \neq r) \\ t_{rj'}^{(L)} - t_{j'}^{(L)} / t_{rj}^{(L)} / w_{rr}^{(L)} & (j \neq r, j' \neq r) \\ 1 / t_{rr}^{(L)} & (j = r, j' = r) \\ - t_{jr}^{(L)} / t_{rr}^{(L)} & (j \neq r, j' = r) \end{cases} \tag{4-119}$$

◆思考和练习题◆

1.什么是趋势分析?试述其计算步骤。
2.趋势面分析在石油和天然气地质中有哪些应用?在应用时应注意哪些问题?
3.为什么作聚类分析时要进行原始数据变换?常用的变换有哪几种?
4.常用的相似性统计量有哪些?它们如何刻画变量和样品间的亲疏关系?
5.主成分分析的几何意义是什么?写出主成分分析的计算步骤。
6.常用的判别分析有哪几类?叙述贝叶斯判别准则的基本思想。

第五章 随机函数及其结构分析

本章学习目的和要求：理解随机函数、随机过程和区域化变量概念的内涵，掌握结构分析的工具——相关图和变差函数及变差函数的计算、拟合和套合，为第六和第七章的学习奠定理论基础。本章内容的学习需以第二章的内容为基础。

第一节 随机函数的定义及其研究意义

一、随机函数的定义

在第二章中，我们介绍了随机变量和 n 维随机向量的概念，当把 n 维随机向量理解为“同一属性的变量在不同点上”时，n 维随机向量就是一个随机函数，可称作空间和/或时间上相关的一组随机变量。

随机函数在一维时间轴上称为随机过程，记为 $Z(t)$，t 为时间轴；在空间域中称为随机场，记为 $Z(x)$，x 为空间位置。若把时间和空间问题结合起来，就构成一个时空模型，记为 $Z(x,t)$。

电场、磁场、重力场和温度场等都是空间点的函数，且它们也都有随机性的变化，故都可看成随机场。对随机场每次观测的结果，都得到一个确定的空间点函数（即实现，realization），而当我们对空间某个点 x_0 来考察随机场时，则得到一个随机变量 $Z(x_0)$。

以上是关于随机函数的直观理解，下面我们从随机试验出发，来定义随机函数：

设随机试验 E 的样本空间为 $\Omega=\{\omega\}$，若对每一个 $\omega\in\Omega$，都有一个函数 $Z(x_1,x_2,\cdots,x_n,\omega)$（$x_1\in X_1,x_2\in X_2,\cdots,x_n\in X_n$）与之对应，且当各自变量 $X_1,X_2,\cdots,X_n$ 均取任一固定值时，函数 $Z(x_1,x_2,\cdots,x_n,\omega)$ 为一随机变量，则称 $Z(x_1,x_2,\cdots,x_n)$ 为定义在 $\{X_1,X_2,\cdots,X_n\}$ 上的随机函数，记为 $Z(x)$。

根据以上定义来理解，随机函数 $Z(x)$ 既可看作是无穷多个实现的组合；也可看成在时间上或空间上相关的一组随机变量，x 代表空间位置。

二、随机函数研究的意义

经典统计学方法是基于以下几个基本假设进行各种分析的：①变量与空间和/或时间无关；②研究对象是纯随机的；③可以无限重复随机试验，进行大量观测；④每次抽样都是独立的。

从经典统计学方法的几个假设来看，其适用范围主要限制在对人为随机试验结果的分析上。然而，地质变量是在空间和/或时间上展布的变量，这些变量在空间和/或时间上有一定的位置，一般来说具有相关性或结构性，而不是纯随机的；另外，对一个样品点取样，只能取一次，而不能进行无限重复取样。在地质变量研究中直接引入经典的概率统计方法，就会遇到困难。

传统的概率论、数理统计及多元统计学方法不能考虑空间和/或时间上展布的变量中的结构性信息，从而难以合理地表征这些在空间和/或时间上具有结构性的变量。

20世纪60年代以来，随着人类对各种自然和社会现象随时间演化规律的兴趣和社会发

展对预测的需要，时间序列分析预测方法得到了长足的发展，该方法旨在根据变量过去的观测值，利用其中所包含的结构性信息建立模型，以达到对变量未来可能出现的值进行预测的目的，在众多领域中，尤其在经济学研究的领域中，该方法得到了广泛的应用。

20世纪70年代诞生的《地质统计学》，可以更好地反映变量的空间结构性和变量值的变化。《地质统计学》已成为以变差函数为基本工具，在研究区域化变量的空间分布结构特征规律性的基础上，选择各种合适的克立金法，以达到更精确地估计或对区域化变量进行随机模拟为主要目的的一门数学地质独立分支学科。

事实上，在时间序列分析预测中的随机过程，就是在一维时间轴上展布的随机函数，而在地质统计学中的区域化变量就是在一维、二维或三维空间域中展布的随机函数。故在对空间和/或时间上展布的变量进行研究时，有必要引进随机函数的概念。

第二节　随机过程及其结构性分析

一、随机过程

1.随机过程的定义

在一维时间轴上展布的随机函数称为随机过程，记为 $Z(t)$，t 是时间。对某一固定的时刻 t_i，$Z(t_i)$是一个随机变量，随机过程 $Z(t)$可看作是无穷多个随机变量 $Z(t_i)$的集合，t 覆盖无限长的时间域(图 5-1)。根据随机过程中时间参量的连续或离散性质，可将随机过程分为连续型随机过程和离散型随机过程。若随机过程 $Z(t)$对于任何的 $t_i \in T$，$Z(t_i)$是一个连续型随机变量，则称此随机过程为连续型随机过程。若随机过程 $Z(t)$对于任何的 $t_i \in T$，$Z(t_i)$是一个离散型随机变量，则称此随机过程为离散型随机过程。若随机过程 $Z(t)$的时间变化范围 T 是有限或无限区间，则称 $Z(t)$是连续参数随机过程。若随机过程 $Z(t)$的时间是有限个或可列个数的集合，则称 $Z(t)$是离散参数随机过程或称为随机变量序列。

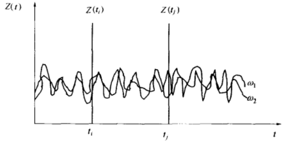


图 5-1　随机过程图示

图5-1中 ω_1 和 ω_2 为随机过程的两个轨迹(或实现)，该连续型随机过程也可看作是无穷多个实现的集合。

2.随机过程的数字特征

若 $E[|Z(t)|^2] < \infty (\forall t)$，则 $Z(t)$称为二阶随机过程，此时自然有 $E[|Z(t)|] < \infty (\forall t)$。

二阶随机过程的一阶原点矩定义为随机变量 $Z(t_i)$ 的数学期望,即

$$\mu_Z(t_i) = E[Z(t_i)] \qquad t_i \in t \tag{5-1a}$$

注意,$E[Z(t_i)]$ 是所有样品函数在时刻 t_i 的函数值的平均,通常称这种平均为集平均,以区别于时间平均。$\mu_Z(t_i)$ 表示了随机过程 $Z(t)$ 在时刻 t_i 的摆动中心。

二阶随机过程的二阶原点矩定义为随机变量 $Z(t_i)$ 的均方值,即

$$\Psi_Z(t_i) = E[Z^2(t_i)] \qquad t_i \in t \tag{5-1b}$$

二阶随机过程的二阶中心矩定义为随机变量 $Z(t_i)$ 的方差 $\sigma_z^2(t_i)$,$\sigma_z(t_i)$ 称为均方差,且有

$$\sigma_z^2(t_i) = D[Z(t_i)] = E\{[Z(t_i) - \mu_z(t_i)]^2\} \qquad t_i \in t \tag{5-1c}$$

对于随机过程在两个不同时刻状态之间的联系,需在二维概率密度的基础上,定义新的数字特征量。设 t_i 和 t_j 是随机过程的任意两个时刻,$Z(t_i)$ 和 $Z(t_j)$ 是该两个时刻的状态,则二阶随机过程的二阶原点混合矩定义为

$$S\{Z(t_i),Z(t_j)\} = E[Z(t_i)Z(t_j)] \qquad (t_i \in t, t_j \in t \;\; i \neq j) \tag{5-2a}$$

类似地,二阶随机过程的二阶中心混合矩定义为

$$\begin{aligned} Cov\{Z(t_i),Z(t_j)\} &= C_{zz}(t_i,t_j) = E\{[Z(t_i) - EZ(t_i)][Z(t_j) - EZ(t_j)]\} \\ &= E[Z(t_i)Z(t_j)] - E[Z(t_i)]E[Z(t_j)] \qquad (t_i \in t, t_j \in t \;\; i \neq j) \end{aligned} \tag{5-2b}$$

二阶中心混合矩也称为随机过程 $Z(t)$ 的自协方差函数,或两个随机变量 $Z(t_i)$ 和 $Z(t_j)$ 之间的协方差函数,记为 C_{zz}。

此处,协方差函数的定义与经典统计学中协方差函数的定义从形式上看是完全相同的[见式(2-10)],但其物理含义有所不同,经典统计学中的协方差函数反映的是两个随机变量之间的相关性,而此处的协方差函数反映的是同一随机过程在不同时间点上的两个随机变量之间的相关性,它反映随机过程的自相关性。

二、平稳随机过程

平稳随机过程,是统计特征不随时间的平移而变化或说不随时间原点的选取而变化的随机过程。用严格的数学术语表达即是:对时间 t 的任意 n 个数值 $t_1, t_2, \cdots, t_n$ 和任意实数 ε,随机过程 $Z(t)$ 的 n 维概率密度函数满足关系

$$f(z_1,z_2,\cdots z_n;t_1,t_2,\cdots t_n) = f(z_1,z_2,\cdots z_n;t_1+\varepsilon,t_2+\varepsilon,\cdots t_n+\varepsilon) \tag{5-3a}$$

则称随机过程 $Z(t)$ 为平稳随机过程,简称平稳过程。

利用上述随机过程数字特征的各个量的定义,容易验证对二阶平稳随机过程有

$$\begin{cases} E[Z(t)] = \mu & \forall t \\ C_{ZZ}(t_i,t_j) = E[Z(t_i)Z(t_j)] - \mu^2 = C_{ZZ}(\tau) & \tau = t_i - t_j \end{cases} \tag{5-3b}$$

事实上,对一阶平稳随机过程的一维概率密度函数,只要令 $\varepsilon = -t_i$,即有

$$f(z_i;t_i) = f(z_i;t_i+\varepsilon) = f(z_i;0) \tag{5-3c}$$

这表明,平稳随机过程的一维概率密度不依赖于时间,记为 $f(z_i)$,于是 $Z(t)$ 的均值为常数,记为 μ,且有

$$E[Z(t_i)] = \int_{-\infty}^{\infty} z_i f(z_i)\,\mathrm{d}z_i = \mu \tag{5-3d}$$

同理，也可以证明，对二阶平稳随机过程，其均方值和方差也为常数，请读者自行验证。

对二阶平稳随机过程的二维概率密度函数，同样令 $\varepsilon = -t_i$ 和 $\tau = t_j - t_i$，即有

$$f(z_i, z_j; t_i, t_j) = f(z_i, z_j; t_i + \varepsilon, t_j + \varepsilon) = f(z_i, z_j; 0, \tau) \tag{5-4a}$$

这个关系式表明，二阶平稳随机过程的二维概率密度只依赖于时间间隔 τ，而与时间的取值 t_i, t_j 无关，记为 $f(z_i, z_j; \tau)$。因此，若设 $t = t_i, \tau = t_j - t_i$，则随机过程 $Z(t)$ 的二阶混合矩可表示为

$$S_{zz} = E[Z(t)Z(t+\tau)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_i z_j f(z_i, z_j; \tau) \mathrm{d}z_i \mathrm{d}z_j \tag{5-4b}$$

随机过程 $Z(t)$ 的协方差函数 $C_{zz}(\tau)$ 可以表示为

$$C_{zz}(\tau) = E\{[Z(t) - \mu][Z(t+\tau) - \mu]\} = S_{zz}(\tau) - \mu^2 \tag{5-4c}$$

由以上的分析知道，二阶平稳随机过程的均值为常数，二阶混合矩 S_{zz} 为单变量 τ 的函数，而且其协方差函数具有以下重要特性：

(1) $C_{ZZ}(\tau)$ 是一个相对于纵轴对称的函数，即

$$C_{ZZ}(\tau) = C_{ZZ}(-\tau) \tag{5-5a}$$

(2) 当 $\tau = 0$ 时，协方差即为变量的方差，变量的方差总大于 0，即

$$C_{ZZ}(0) = \sigma^2 > 0 \tag{5-5b}$$

(3) 满足 Schwartz 不等式，即

$$C_{ZZ}(\tau) \leqslant C_{ZZ}(0) \tag{5-5c}$$

式(5-5c)表明，协方差函数是一个递减函数，其物理含义是指随两点间时间间隔的增大，随机过程的自相关性减弱。

若对于给定的随机过程 $Z(t)$，有 $E[Z(t)] = \mu$ 为常数，而且

$$E[Z^2(t)] < +\infty, E[Z(t)Z(t+\tau)] = S_{zz}(\tau) \tag{5-6}$$

则称随机过程 $Z(t)$ 为宽平稳过程或广义平稳过程。

三、二阶平稳假设条件下的自相关函数及其性质

同协方差函数一样，自相关函数也是表征随机过程时间相关结构性的一个重要函数，与协方差函数有如下关系

$$\rho_{ZZ}(\tau) = \frac{C_{ZZ}(\tau)}{C_{ZZ}(0)} \tag{5-7}$$

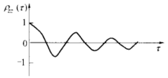


图 5-2 一个随机过程的自相关图

$\rho_{ZZ}(\tau)$ 与 τ 的关系图称为自相关图（图 5-2）。

自相关函数具有以下 4 个性质：

(1) $\rho_{ZZ}(0) \equiv 1$ 且 $|\rho_{ZZ}(\tau)| \leqslant 1 (\tau \neq 0)$；

(2) $\rho_{ZZ}(\tau)$ 的周期与随机过程的周期相同；

(3) 当 $\rho_{ZZ}(\tau)$ 很快趋近于 0 时，随机过程为平稳过程，反之，则为非平稳过程；

(4) 对于白噪音，$\rho_{ZZ}(0) = 1$，$\rho_{ZZ}(\tau) = 0 (\tau \neq 0)$，白噪音是无时间相关结构的一种特殊随机过程，是一种纯粹随机的过程。

第三节　区域化变量及其变差函数的功能

一、区域化变量的涵义

1.区域化变量的定义

地质变量多为在空间上展布的变量,地质变量具有区域性特点,即在一定的区域内展布,在地质统计学中,把地质变量称作区域化变量,如储集层的孔、渗、饱及盖层厚度等,都是区域化变量。

区域化变量是在空间里的随机函数,在空间上的任一点,都对应有一个随机变量,不同空间点上的随机变量之间具有相关性。所以,区域化变量既能反映空间变量的结构性,又能反映其随机性,具有双重性的功能,这是地质统计学与经典统计学的本质区别所在。

区域化变量在观测前为随机场,在观测后为三元实值函数或空间点函数,凡是在空间上展布的地质变量都可称为区域化变量。

2.区域化变量的物理学或地质学特性

1)区域化变量有明确的物理含义

如孔隙度、渗透率等地质变量的物理含义是明确的,是随机函数在研究空间变量中的具体应用,而随机函数本身只具有纯粹的数学意义。

2)空间局限性

区域化变量往往只存在于一定的空间范围内,如含油饱和度只限于油藏范围内,这一空间称为区域化变量的几何域。空间局限性还表现在,区域化变量是按几何承载来确定的,如孔隙度是在一定直径的岩心上确定的,如果承载变了则会得到一个新的区域化变量。

3)不同程度的连续性

不同的区域化变量具有不同程度的连续性,也就是说不同的区域化变量的空间结构性或相关性是不同的。

4)不同类型的各向异性

区域化变量在各个方向上如果性质相同,则称其为各向同性的,否则,称其为各向异性的。地质变量往往是各向异性的,且各向异性的类型也有不同。

二、区域化变量的协方差函数和变差函数

在第二节中,通过随机过程的概念,引入了与时间变量或时间间隔有关的协方差函数,对区域化变量的随机函数,只需将上述时间变量替换成空间位置变量,那么上述各个数字特征量的定义及性质可以同样地成立。

不妨设 $Z(x)$和 $Z(x+h)$为相距 h 的两个随机变量(x 和 $x+h$ 是空间上相距为 h 的两点),则这两个随机变量之间的协方差函数为

$$\begin{aligned} Cov[Z(x),Z(x+h)] &= C(x,x+h) \\ &= E\{Z(x)-E[Z(x)]\}\{Z(x+h)-E[Z(x+h)]\} \\ &= E[Z(x)Z(x+h)]-E[Z(x)]E[Z(x+h)] \quad \forall x \quad \forall h \end{aligned} \tag{5-8}$$

由此可见,协方差函数依赖于空间点 x 和h,当 $h=0$ 时有

$$C(x,x) = E[Z(x)^2] - \{E[Z(x)]\}^2 = \mathrm{Var}[Z(x)] \quad \forall x \tag{5-9}$$

式(5-9)中 $\mathrm{Var}[Z(x)]$ 为 x 点处随机变量 $Z(x)$ 的方差或称验前方差。

$Z(x)$ 和 $Z(x+h)$ 这两个随机变量之间的协相关系数为

$$\rho(x,x+h) = \frac{\mathrm{Cov}[Z(x),Z(x+h)]}{\sqrt{\mathrm{Var}[Z(x)]\mathrm{Var}[Z(x+h)]}} \quad \forall x, \quad \forall h \tag{5-10}$$

在地质统计学中,为考察区域化变量的空间变化特征,引入变差函数的概念,变差函数定义为增量方差之半,即

$$\gamma(x,x+h) = \frac{1}{2}\mathrm{Var}[Z(x) - Z(x+h)] \tag{5-11}$$

$$= \frac{1}{2}E[Z(x) - Z(x+h)]^2 - \frac{1}{2}\{E[Z(x)] - E[Z(x+h)]\}^2 \quad \forall x, \forall h$$

三、随机函数的平稳假设和本征假设

1.假设的必要性

要估计变差函数值,就要能估计数学期望值 $E[Z(x)-Z(x+h)]^2$、$E[Z(x)]$ 和 $E[Z(x+h)]$[见式(5-11)],这就必须有若干对 $Z(x)$ 和 $Z(x+h)$ 的值。实际工作中这一条件是不可能满足的,因为地质问题不同于人为的随机试验,不可能在一个点上重复取样,无法对 $E[Z(x)-Z(x+h)]^2$、$E[Z(x)]$ 和 $E[Z(x+h)]$ 进行估计,这就在统计推断上发生了困难。为了克服这一困难,就需要对 $Z(x)$ 做一些假设。

2.平稳假设

平稳假设又分为强平稳假设和弱平稳假设。

1)强平稳假设

假设区域化变量 $Z(x)$ 的任意 n 维分布函数(联合分布)均不因空间点 x 发生位移 h 而改变,则称区域化变量为强平稳,用公式表达为

$$F_{x_1,x_2,\cdots,x_n}(z_1,z_2,\cdots,z_n) = P\{Z(x_1) \leqslant z_1, Z(x_2) \leqslant z_2, \cdots, Z(x_n) \leqslant z_n\}$$

$$= P\{Z(x_1+h) \leqslant z_1, Z(x_2+h) \leqslant z_2, \cdots, Z(x_n+h) \leqslant z_n\}$$

$$= F_{x_1+h,x_2+h,\cdots,x_n+h}(z_1,z_2,\cdots,z_n) \quad \forall n, \forall h, \forall x_1,x_2,\cdots,x_n \tag{5-12}$$

2)弱平稳假设

强平稳假设只有数学上的意义,在实际应用中,很难找到满足强平稳假设的变量,这就需要我们放宽假设条件,以满足实际中遇到的大多数变量,一般采用弱平稳假设,即只假设 $Z(x)$ 的一阶矩和二阶矩存在且平稳,称为二阶平稳,即

$$\begin{cases} E[Z(x)] = \mu \quad \forall x \\ \mathrm{Cov}[Z(x),Z(x+h)] = C(h) = E[Z(x)Z(x+h)] - E[Z(x)]E[Z(x+h)] \\ = E[Z(x)Z(x+h)] - \mu^2 \quad \forall x, \forall h \end{cases} \tag{5-13}$$

在二阶平稳假设的条件下,变差函数式(5-11)可写为

$$\gamma(x,h) = \gamma(h) = \frac{1}{2}E[Z(x) - Z(x+h)]^2 \quad \forall x, \forall h \tag{5-14}$$

有了二阶平稳假设,我们把空间上相距为 h 的所有数据点找出来构成数据点对,利用这些数据点对即可对 $E[Z(x)-Z(x+h)]^2$、$E[Z(x)]$ 和 $E[Z(x+h)]$ 进行估计,这就解决了上

述统计推断的困难。

在平稳假设条件下，变差函数和协方差函数之间存在如下关系

$$\gamma(h) = C(0) - C(h) \tag{5-15}$$

图 5-3 展示了 $\gamma(h)$和 $C(h)$之间的对应关系。

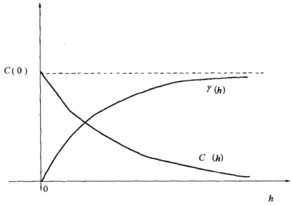


图 5-3　$\gamma(h)$与 $C(h)$之间的关系示意图

注：式(5-15)的推导过程如下：

$$\begin{aligned}\gamma(h) &= \frac{1}{2}E[Z(x) - Z(x+h)]^2 = \frac{1}{2}E\{[Z(x) - \mu] - [Z(x+h) - \mu]\}^2 \\ &= \frac{1}{2}E[Z(x) - \mu]^2 - E[Z(x) - \mu][Z(x+h) - \mu] + \frac{1}{2}E[Z(x+h) - \mu]^2 \\ &= \frac{1}{2}C(0) - E[Z(x) - \mu][Z(x+h) - \mu] + \frac{1}{2}C(0) \\ &= C(0) - E[Z(x) - \mu][Z(x+h) - \mu] = C(0) - C(h)\end{aligned}$$

式(5-15)推导完毕。

3. 本征假设

为了更加弱化假设条件以便适应更多的地质变量，引入本征假设，该假设要求增量方差存在且平稳，即要求 $Z(x)$满足以下条件：

$$\begin{cases} E[Z(x) - Z(x+h)] = 0 \quad \forall x, \forall h \\ Var[Z(x) - Z(x+h)] = E[Z(x) - Z(x+h)]^2 - \{E[Z(x) - Z(x+h)]\}^2 \\ = E[Z(x) - Z(x+h)]^2 \\ = 2\gamma(h) \quad \forall x, \forall h \end{cases} \tag{5-16}$$

4. 本征假设与二阶平稳假设的比较

本征假设与二阶平稳假设相比，其条件更弱，适用的变量更多，满足二阶平稳假设的变量一定满足本征假设，而满足本征假设的变量不一定满足二阶平稳假设，下面给出的例子就是一个满足本征假设而不满足二阶平稳假设的情况。

【例 5-1】　变量 Z_n 由以下方式构成

$$Z_n = Z(n) = \sum_{i=1}^{n} X_i, X_i = \begin{cases} 1 & P = 1/2 \\ -1 & P = 1/2 \end{cases} \quad (X_i \text{ 相互独立}, i = 1, 2, \cdots, n)$$

问 Z_n 是否满足平稳假设？是否满足本征假设？

解：Z_n 的期望值为 $E[Z_n] = \sum_{i=1}^{n} E[X_i] = 0$

Z_n 的方差为

$$Var(Z_n) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \{EX_i^2 - [E(X_i)]^2\} = \sum_{i=1}^{n} E(X_i^2) = n$$

当 $n \to \infty$ 时，Z_n 的方差也趋近于无穷大，故 Z_n 的方差和协方差不存在，所以 Z_n 不满足平稳假设，然而 Z_n 的变差函数

$$\gamma(h) = \gamma(m-n) = \frac{1}{2}E[Z(m) - Z(n)]^2 = \frac{1}{2}E[\sum_{i=1}^{m} X_i - \sum_{i=1}^{n} X_i]^2 = \frac{|m-n|}{2} = \frac{h}{2}$$

是存在且平稳的。

因此，Z_n 满足本征假设，但不满足二阶平稳假设。

四、在二阶平稳假设条件下变差函数的性质和功能

在二阶平稳假设条件下，区域化变量的协方差函数 $C(h)$ 与随机过程的协方差函数 $C_{ZZ}(\tau)$ 具有相似的性质，在此不再赘述。惟一不同的是 $C(h)$ 必须是非负定的函数，即要求由 $C(x_i, x_j)$ 构成的协方差函数矩阵必须是非负定矩阵，这一要求是为了保证后面的克立金方程组（见第六章）有且只有一组解，下面仅就变差函数的性质和功能进行讨论。

1. 变差函数的性质

(1) 协方差函数反映变量的相关性，而变差函数反映变量的变化性，二者皆为区域化变量结构性表征的统计工具，二者呈互为消长的关系，$\gamma(h)$ 与 $C(h)$ 之间存在式(5-15)所示的关系（见图 5-3）；

(2) 当 h 为 0 时，即变量与其本身相比没有发生变化，故有 $\gamma(0) = 0$；

(3) 根据变差函数的定义[式(5-11)或(5-14)]，变差函数是非负函数，即 $\gamma(h) \geqslant 0$；

(4) 变差函数是一个相对于纵轴对称的函数，即 $\gamma(h) = \gamma(-h)$；

(5) $-\gamma(h)$ 必须是条件非负定函数，即要求由 $-\gamma(x_i, x_j)$ 构成的矩阵当 $\sum_{i=1}^{n} \lambda_i = 0$ 时为非负定矩阵，这一要求是为了保证后面的克立金方程组（见第六章）有且只有一组解；

(6) 当 h 趋近于无穷大时，变差函数达到最大值，该最大值为变量的方差，即 $\gamma(\infty) = C(0)$。

2. 变差函数的功能

一个典型的变差函数具有如图 5-4 所示的形态，图中，$C(0)$ 为变量的验前方差，在地质统计学中称为基台值，C_0 为块金值，C 为拱高，a 为变程，变差函数有以下几个功能：

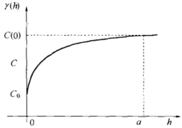


图 5-4　一个典型的变差函数图

(1) 通过变程反映变量的影响范围（如储集体的分布范围）；

(2) 变差函数在原点处的性状可以反映变量的空间连续性，图 5-5 中，从抛物线型到直线型到间断型再到随机型，变量的连续性逐渐变差；

(3) 不同方向上的变差函数图可反映区域化变量的各向

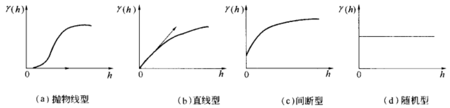


图5-5 变差函数在原点处的性状反映变量的空间连续性

异性；

(4)变差函数如果是跃迁型的，其基台值的大小可反映变量变化幅度的大小；

(5)块金常数 C_0 的大小可反映区域化变量随机性的大小，反映区域化变量的微观变化和测量误差；

(6)进行变量的结构性分析，研究变差函数，是进行克立金估计和随机模拟的基础，变差函数给出了变量的结构性信息，是克立金估计和随机模拟所需要的，正是利用了变量的结构性信息，克立金估计方法与传统的估计方法相比，才具有了其优越性。

第四节　实验变差函数的计算和拟合

一、实验变差函数的计算

根据变差函数的定义 $\gamma(h)=\frac{1}{2}E[Z(x)-Z(x+h)]^2$ 可知，实验变差函数的计算公式为

$$\gamma^*(h)=\frac{1}{2N(h)}\sum_{k=1}^{N(h)}[Z(x_k)-Z(x_k+h)]^2 \qquad (5-17)$$

其中，$N(h)$为相距为 h 的数据点对数目。

1.规则和基本规则取样条件下实验变差函数的计算

在规则或基本规则取样条件下，计算变差函数比较简单，在如图5-6所示的正方形取样网格中，设正方形的边长为 ℓ，各节点上的数据代表变量在节点处的取值。

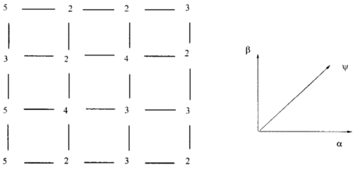


图5-6 规则取样图示(节点上的数值代表变量值)

在规则网格情况下，基本步长一般取网格的长度（或宽度），步长取基本步长的倍数，如在 α 方向上，$h=3\ell$ 时的实验变差函数值为

$$\gamma_\alpha^*(3\ell)=\frac{1}{2}\times\frac{1}{4}(3^2+2^2+1^2+2^2)=\frac{1}{2}\times\frac{1}{4}\times 18=\frac{9}{4}=2.25$$

同理可求函数 $\gamma_\alpha^*(\ell)$、$\gamma_\alpha^*(2\ell)$、$\gamma_\beta^*(3\ell)$和 $\gamma_\psi^*(\sqrt{2}\ell)$等。

2.不规则取样条件下实验变差函数的计算

数据点不规则分布是最常见的现象，在计算变差函数时，按规则网格情况下的计算方法就会行不通，这时，要给出角度容差 d 和距离容差 ε，落在图 5－7 所示扇形区域中的点就认为是在 α 方向上与 x_0 点相距为 h 的点，据此，确定了在 α 方向上与 x_0 点相距为 h 的数据点对后，就可根据式（5－17）计算实验变差函数。

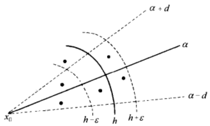


图 5－7　不规则网格条件下实验变差函数计算时数据点对的确定方法

d 为角度容差，ε 为距离容差

3.变差函数计算的一般步骤

由于区域化变量常常存在各向异性，即不同方向上的变差函数不同，变差函数的一个重要功能就是要反映区域化变量的各向异性特征，实验变差函数一般按以下步骤计算：

（1）从单向性变差函数的计算开始，注意这样做并不意味着假设变量是各向同性的。这样做的目的，一是作为计算变差函数的开端，对确定变差函数计算中的某些参数有用，二是为了避免在计算方向性变差函数时的某些偏差。

（2）在单向性变差函数计算的基础上，计算方向性变差函数。在许多实际问题中，常常有些已知的信息（如河道砂体展布方向等信息）表明各向异性的方向，这些信息有助于方向性变差函数计算时方向的选择。若事先不知道各向异性的方向，这时，在计算方向性变差函数时，就要通过不同方向上的变差函数的玫瑰花图确定各向异性的方向。

（3）若通过分析证明变量是各向同性的，则此时可求出平均变差函数

$$\gamma^*(h)=\frac{\sum_{i=1}^{n}N_i\gamma_i^*(h)}{\sum_{i=1}^{n}N_i}\qquad(5-18)$$

式中，$\gamma_i^*(h)$为 i 方向上的变差函数；N_i 为 i 方向上变差函数计算时相距 h 的数据点对的个数。

二、实验变差函数的拟合

1.拟合的必要性

1）消除实验变差函数中的不稳定因素

在实验变差函数的计算中，由于数据中存在的随机性、波动性和数据数量的限制（实际应用中，常常出现数据稀少的情况），导致实验变差值计算中存在不确定性和波动性，需要通过拟合来消除这种不确定性和波动性。

2）满足克立金方程组求解的要求

在后面的克立金方法中，建立和求解克立金方程组，需要知道任一方向、任一步长 h 所对应的变差函数值，通过实验变差函数的拟合，得到一个随 h 变化的连续函数，以满足克立金方程组求解的要求。

3)保证克立金方程组有且只有一组解

为了使克立金方程组有而且只有惟一的解，不是任何函数都可作为变差函数来使用的，地质统计学家经过研究和数学论证，已提出了若干函数形式(见后面的基本模型)，这些函数形式是可以作为变差函数来使用的，利用这些函数形式，来拟合实验变差函数，就能够保证克立金方程组有且只有一组解(已经过理论统计学家们的证明)。在具体应用时，采用这些函数形式就足够了，没有必要再去"创造"其它函数形式。

2.拟合的3个条件

1)稳健估计

首先，实验变差函数的估计应该是稳健的，如遇数据中存在特异值的情况，在进行实验变差函数计算时，要剔除这些特异值，以保证实验变差函数估计的稳健性。

2)正确的模型

针对不同的变量及其实验变差函数的形态特点，选取合适的模型进行拟合，才能得到反映变量结构性特点的正确模型。

3)好的拟合方法

可调整模型参数的计算程序，以便于找到最好的拟合。

3.基本模型

变差函数有多种模型，包括有基台值的模型、无基台值的模型、纯块金效应模型和空穴效应模型等。无基台值的模型包括幂函数模型和对数函数模型，纯块金效应模型反映变量是纯随机的，空穴效应模型往往是以某种套合的方式伴随着其它有基台值或无基台值模型而出现。

最常用的模型是有基台值模型，有基台值模型又包括球状模型、指数模型和高斯模型(图5-8)，下面是3种常用的有基台值模型的表达式。

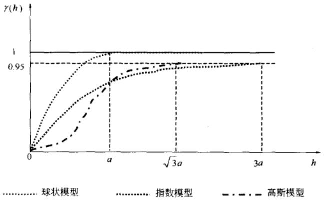


图5-8　三种有基台值模型的形态曲线

1)球状模型

$$\gamma(h)=\begin{cases}0 & (h=0)\\ C_0+C(\frac{3}{2}\frac{h}{a}-\frac{1}{2}\frac{h^3}{a^3}) & (0<h\leqslant a)\\ C_0+C & (h>a)\end{cases} \tag{5-19}$$

2)指数模型

$$\gamma(h)=\begin{cases}0 & (h=0)\\ C_0+C(1-e^{-h/a}) & (h>0)\end{cases} \tag{5-20}$$

3)高斯模型

$$\gamma(h)=\begin{cases}0 & (h=0)\\ C_0+C(1-e^{-h^2/a^2}) & (h>0)\end{cases} \tag{5-21}$$

对球状模型,变差函数达到基台值时对应的 h 值 a 即为变量的变程,对指数模型和高斯模型,变差函数值达到基台值的95%时对应的 h 值分别为 $3a$ 和$\sqrt{3}a$,$3a$ 和$\sqrt{3}a$ 分别可近似地看作指数模型和高斯模型的变程。

三、变差函数实例分析

用变差函数对塔里木盆地轮南凸起的实测地层压力这一区域化变量进行了分析,在进行结构分析之前,我们首先在纵向上进行了压力系统的划分,分为上部常压系统、中部超压系统和深部常压系统。

根据尝试对比,我们选择球状模型作为研究区地层压力的变差函数模型,表5-1列出了塔北隆起3个压力系统地层压力在不同方向上的变差函数模型的参数。

由表5-1可以看出在不同方向上,变程、基台值、块金值差别比较大,特别在中部的超压压力系统内,在东西方向和南北方向上,变程和块金值的差别更为明显(图5-9),反映出地层压力具有明显的各向异性。

表5-1 塔里木盆地轮南凸起不同压力系统的变差函数拟合参数

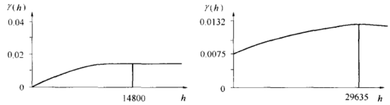
参数 / 压力系统	变程(a),m		基台值($C+C_0$)		块金值(C_0)	
	东西方向	南北方向	东西方向	南北方向	东西方向	南北方向
上部静水压力系统	24300	30000	0.0071	0.0042	0.0015	0.0004
中部超压压力系统	14800	29635	0.0183	0.0139	0.0006	0.0081
下部静水压力系统	27200	30100	0.0091	0.0013	0.0002	0.0000

通过表5-1的参数,可以描述塔北隆起不同压力系统内地层的结构特征:

(1)上部和下部静水压力系统内压力系数的块金值都比较小,反映了地层压力在这两个压力系统内的连续性较好;在超压压力系统内,块金值比较大,反映了超压压力系统内地层压力的随机性比较明显。

(2)3个压力系统内的变差函数都不尽相同,不同方向上变差函数的差异也比较明显,表明塔北隆起内的地层压力呈现较强的各向异性。

(3)上部和下部静水压力系统内压力系数在不同方向上的变程一般在24~30km内,相差



(a) 东西方向的变差函数　　　　(b) 南北方向的变差函数

图 5-9　中部超压压力系统地层压力在东西和南北两个方向上变差函数的拟合图

不大,表明这两个压力系统的地质体结构呈现圆～近圆状;而中部的超压压力系统在南北方向与东西方向上的各向异性比达到 2.002,呈现带状各向异性。

(4)在上述 3 个压力系统内,从总体上看来,东西方向的基台值比南北方向上的要大,反映了地层压力在东西方向上的变化幅度更大。

第五节　实验变差函数的结构套合

实际区域化变量的变化性是很复杂的,往往包含有各种尺度上的多层次的变化性,反映在变差函数上,就是其结构性往往不是单一结构,而是多层次结构叠加在一起——套合结构。大尺度的变化总是包含着小尺度的变化在内的,但是,在只有大尺度观测数据的条件下,不能从大尺度的变化性中区别小尺度的变化性,观测得细,结论也细,观测得粗,结论也就粗。

套合的目的是获得一个反映随着距离和方向变化的变差函数。

一、各向同性条件下的结构套合

各向同性是指区域化变量在不同方向上的变化性完全相同。在此情况下,可直接进行变差函数的套合(叠加),即 $\gamma(h)=\gamma_0(h)+\gamma_1(h)+\cdots+\gamma_i(h)+\cdots$。不同尺度上的结构可以是相同的模型,也可以是不同的模型。如某各向同性区域化变量的变差函数由 $\gamma_0(h)$、$\gamma_1(h)$ 和 $\gamma_2(h)$ 3 个不同尺度上的结构组成,$\gamma_0(h)$、$\gamma_1(h)$ 和 $\gamma_2(h)$ 分别为

$$\gamma_0(h)=\begin{cases}0 & (h=0)\\ C_0 & (h>0)\end{cases} \tag{5-22}$$

$$\gamma_1(h)=\begin{cases}C_1\left(\dfrac{3}{2}\dfrac{h}{a_1}-\dfrac{1}{2}\dfrac{h^3}{a_1^3}\right), & (0\leqslant h\leqslant a_1)\\ C_1 & (h>a_1(a_1<a_2))\end{cases} \tag{5-23}$$

$$\gamma_2(h)=\begin{cases}C_2\left(\dfrac{3}{2}\dfrac{h}{a_2}-\dfrac{1}{2}\dfrac{h^3}{a_2^3}\right) & (0\leqslant h\leqslant a_2)\\ C_2 & (h>a_2)\end{cases} \tag{5-24}$$

则该区域化变量的变差函数为 3 个变差函数 $\gamma_0(h)$、$\gamma_1(h)$ 和 $\gamma_2(h)$ 之和,即

$$
\gamma(h)=\begin{cases}0 & (h=0)\\ C_0+\frac{3}{2}\left(\frac{C_1}{a_1}+\frac{C_2}{a_2}\right)h-\frac{1}{2}\left(\frac{C_1}{a_1^3}+\frac{C_2}{a_2^3}\right)h^3 & (0<h\leqslant a_1)\\ C_0+C_1+C_2\left(\frac{3}{2}\frac{h}{a_2}-\frac{1}{2}\frac{h^3}{a_2^3}\right) & (a_1<h\leqslant a_2)\\ C_0+C_1+C_2 & (h>a_2)\end{cases} \quad (5-25)
$$

各向同性条件下变差函数的套合过程如图 5 - 10 所示。

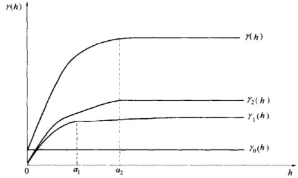


图 5 - 10　各向同性条件下变差函数的叠合过程示意图

二、区域化变量各向异性条件下的结构套合

各向异性指在不同方向上，变量的变化性不同，各向异性又分为几何各向异性、带状各向异性和混合各向异性：在不同方向上基台值相同但变程不同，称为几何各向异性，见图 5 - 11(a)；在不同方向上基台值不同但变程相同，称为带状各向异性，见图 5 - 11(b)；在不同方向上，基台值和变程都不同，称为混合各向异性，见图 5 - 11(c)，混合各向异性是最常见的情况。

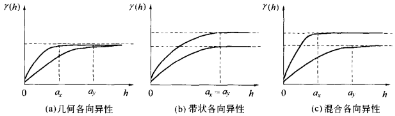


图 5 - 11　不同类型各向异性示意图

1. 几何各向异性情况下的结构套合

当区域化变量为几何各向异性时，要先对同一尺度上的各向异性结构进行各向同性化处理，然后再把同性化后的不同尺度上的结构叠加而实现套合。

1)变差函数的各向同性化变换

假设式(5 - 26)和(5 - 27)分别为某区域化变量在 x 方向和 y 方向上的变差函数

$$\gamma(h_x)=\begin{cases}0 & (h_x=0)\\ C\left[\dfrac{3}{2}\dfrac{h_x}{a_x}-\dfrac{1}{2}\dfrac{h_x^3}{a_x^3}\right] & (0<h_x\leqslant a_x)\\ C & (h_x>a_x)\end{cases} \tag{5 - 26}$$

$$\gamma(h_y)=\begin{cases}0 & (h_y=0)\\ C\left[\dfrac{3}{2}\dfrac{h_y}{a_y}-\dfrac{1}{2}\dfrac{h_y^3}{a_y^3}\right] & (0<h_y\leqslant a_y)\\ C & (h_y>a_y)\end{cases} \tag{5 - 27}$$

这两个变差函数的基台值皆为 $C(0)$,但在两个方向上的变程不同,分别为 a_x 和 a_y,属几何各向异性。

通过式(5 - 28)的坐标变换后(见图 5 - 12),得到一个各向同性化的变差函数[式(5 - 29)]。

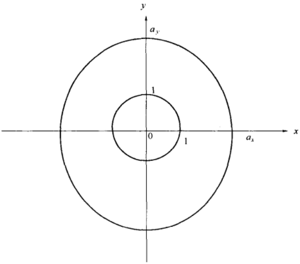


图 5 - 12　几何各向异性各向同性化的坐标变换示意图

$$\vec{h}=\begin{bmatrix}\dfrac{h_x}{a_x}\\ \\ \dfrac{h_y}{a_y}\end{bmatrix},\quad h=\sqrt{(\frac{h_x}{a_x})^2+(\frac{h_y}{a_y})^2} \tag{5 - 28}$$

$$
\gamma(h) = \begin{cases} 0 & (h = 0) \\ C\left[\frac{3}{2}h - \frac{1}{2}h^3\right] & (0 < h \leqslant 1) \\ C & (h > 1) \end{cases} \tag{5-29}
$$

注意 $\gamma(h_x)$和 $\gamma(h_y)$的函数类型必须相同，才能进行式(5－28)的坐标变换。

2)不同尺度上结构的套合

在对同一尺度上的几何各向异性进行各向同性化后，对不同尺度上的结构可进行叠加套合，如式(5－30)即为式(5－31)和(5－32)两个各向同性化之后的变差函数的叠加。

$$
\gamma(h) = \gamma_1(h_1) + \gamma_2(h_2) \tag{5-30}
$$

$$
\gamma_1(h_1) = \gamma_1\left(\sqrt{\left(\frac{h_x}{a_{x1}}\right)^2 + \left(\frac{h_y}{a_{y1}}\right)^2}\right) \tag{5-31}
$$

$$
\gamma_2(h_2) = \gamma_2\left(\sqrt{\left(\frac{h_x}{a_{x2}}\right)^2 + \left(\frac{h_y}{a_{y2}}\right)^2}\right) \tag{5-32}
$$

2.带状各向异性情况下的结构套合

带状各向异性的变差函数可以看作一种不同方向结构的套合结构，在这种套合结构中，可直接把各部分相加，即

$$
\gamma(h) = \sum_{i=1}^{n} \gamma_i(h) \tag{5-33}
$$

3.混合各向异性情况下的套合结构

在混合各向异性中，每一个组成结构可以具有各自不同的异向性但都可以通过适当的线性变换化为各向同性后叠加，其中，有的 $\gamma_i(h)$可以是几何异向性的，对每个组成结构分别进行一种适当的坐标线性变换，把它们转化为各向同性结构，最后，再把这些各向同性结构套合(叠加)起来，构成一个统一的变差函数模型。如对图 5－13 所示的情况(a_x、a_y 和代表水平方向上区域化变量的变程，a_z 代表垂直方向上区域化变量的变程)，可以把水平方向各向同性化后，再加上垂直方向的结构。

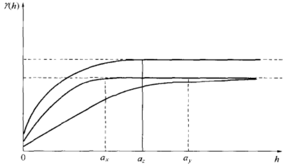


图 5－13　混合各向异性变差函数示意图

水平方向各向同性化后的结构为

$$\gamma_1(h_1) = \gamma_1\left(\sqrt{\left(\frac{h_x}{a_x}\right)^2 + \left(\frac{h_y}{a_y}\right)^2}\right) \tag{5-34}$$

垂直方向结构为

$$\gamma_2(h_2) = \gamma_2\left(\frac{h_z}{a_z}\right) \tag{5-35}$$

套合后的结构为

$$\gamma(h) = \gamma_1(h_1) + \gamma_2(h_2) \tag{5-36}$$

◆ 思 考 和 练 习 题 ◆

1.如何理解随机函数这一概念?
2.什么是区域化变量?它的两重性是什么?
3.何谓平稳假设和本征假设?二者有何区别?为什么要做假设?
3.二阶平稳假设条件下自相关函数有哪些性质?
4.二阶平稳假设条件下变差函数有哪些性质和功能?
5.区域化变量有何地质特性?
6.尝试推导关系式 $\gamma(h) = C(0) - C(h)$。
7.如何理解几何各向异性和带状各向异性?
8.变差函数有哪些基本模型?在各向异性条件下如何进行套合?

第六章　克立金分析方法

本章学习目的和要求：了解线性估计的基本概念，理解普通克立金方法的基本原理和克立金方程的建立过程，掌握指示克立金方法的原理，为随机模拟奠定基础。本章学习需以第五章的内容为基础，同时，第一章至第三章的基础知识也需要具备，尤其是第三章中的最小二乘原理必须掌握。

第一节　估计问题和线性估计方差

一、估计问题和估计的一般形式

1.估计问题

在地质变量的定量化研究中，我们常常碰到的问题是根据实测样品值对未取样点或未取样域的变量值进行估计。例如，在油气勘探中，应用井中数据和其它信息，预测未钻井处地层的各种性质，进而对油藏或储层的各种参数进行预测。

图 6－1 中的样品值 $z_{\nu_1}, z_{\nu_2}, \cdots, z_{\nu_n}$ 是在承载 $\nu_1, \nu_2, \cdots, \nu_n$ 上测得的，$\nu_1, \nu_2, \cdots, \nu_n$ 称为信息域，记为 ν。根据这些样品值对 V 内的变量值进行估计（V 称为待估域），就是我们要讨论的估计问题（图 6－1），这一问题的实质是估计出变量在待估域 V 内的平均值，当然，待估域 V 也可以是无限小的一个域，此时，待估域就变成了待估点。

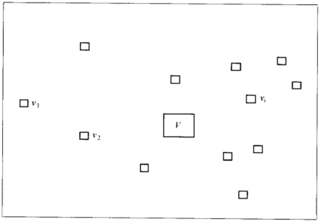


图 6－1　变量估计（内插）的图示

估计方法有很多，如线性内插、滑动平均、三角形加权、多边形加权和简单克立金及普通克立金等方法，其中，简单克立金方法和普通克立金是一种线性的、无偏的和最优的估计。

在图 6－1 中，若待估域 V 非无限小，则用于估值的克立金方法称为块克立金，若待估域 V

无限小，则用于估值的克立金方法称为点克立金，点克立金是块克立金的特例，在本章克立金方法的讨论中，我们以块克立金方程组的推导展开。

另外，我们在实际工作中也可能会遇到这样一类问题：在已知信息域 ν 的条件下，问待估域 V 中变量值大于某个给定值(或称阈值)的概率是多大？这类问题实质上是估计随机变量 $Z(V)$ 的概率分布。这类问题的解决，有广泛的实际应用，如采矿工程中可采储量的估算、环境工程中污染物浓度分布的估计和油气藏工程中高孔渗带的分布等。解决概率分布估计问题常用的方法是指示克立金方法，指示克立金方法是一种非线性的、无偏的和最优的估计，在本章介绍了简单克立金方法和普通克立金方法之后，我们接着就介绍指示克立金方法。

2. 估计的一般形式

由数理统计我们知道，任何估计都会有误差，真实值与估计值之差就是估计误差，在地质统计学中也不例外。

设 $Z(x)$ 为一个二阶平稳区域化变量，即 $Z(x)$ 满足式(5－13)所要求的条件，其协方差函数 $C(h)$ 和变差函数 $\gamma(h)$ 存在且平稳。

区域化变量 $Z(x)$ 在 ν_i 和 V 上的值实际上是 $Z(x)$ 在 ν_i 和 V 上的平均值，即

$$Z(\nu_i) = \int_{\nu_i} Z(t)\mathrm{d}t \tag{6-1}$$

$$Z(V) = \int_V Z(y)\mathrm{d}y \tag{6-2}$$

若用 $Z^*(V)$ 代表用信息域数据对待估域的估值，则估计误差为

$$R(V) = Z(V) - Z^*(V) \tag{6-3}$$

可以证明，若 $Z(x)$ 是二阶平稳的，则 $R(V)$ 也是二阶平稳的，即

$$E[R(V)] = \mu_E \qquad (\forall x, \mu_E \text{ 一般为 } 0) \tag{6-4}$$

$$Var[R(V)] = \sigma_E^2 \qquad (\forall x) \tag{6-5}$$

这表明，当 $Z(x)$ 为二阶平稳且 $R(V) = Z(V) - Z^*(V)$ 时，$R(V)$ 的均值为0，估计方差 $\sigma_E^2 = E[Z(V) - Z^*(V)]^2$ 总是存在且平稳。

估计方差是衡量一个估计好坏的重要尺度。但在寻求一个好的估计量之前还必须考虑估计量的函数形式。

估计 $Z(V)$ 的问题，最一般的形式是求函数 f，使 $Z^*(V) = f\{z(\nu_1), z(\nu_2), \cdots, z(\nu_n)\}$ 满足以下两个条件：

(1)无偏性—$E[Z^*(V)] = E[Z(V)]$；

(2)最优性—估计方差 $\sigma_E^2 = E[Z(V) - Z^*(V)]^2$ 最小。

为此，首先要计算出 $E[Z(V) - Z^*(V)]$ 和 $E[Z(V) - Z^*(V)]^2$ 来，以便再从 f 要满足的两个条件求得一个方程组，从而解出 f 来。而要计算这两个数学期望，就要知道 $(n+1)$ 维随机变量的联合分布，即采用"条件数学期望"

$$Z^*(V) = E\{Z(V)/z(\nu_1), z(\nu_2), \cdots, z(\nu_n)\} \tag{6-6}$$

实际上，我们只能得到 n 维随机变量的一个实现，故无法知道上述的联合分布，采用最一般的函数形式 f 进行估计遇到困难，下面，我们探讨用线性估计形式来解决估计问题的可能性。

3. 线性估计形式

如果我们限制函数 f，只取线性函数形式的话，就可以根据 $Z(x)$ 的协方差函数(或变差函

数)来计算出 $E[Z(V)-Z^*(V)]$和 $E[Z(V)-Z^*(V)]^2$,因此,我们将采取线性估计形式

$$Z^*(V)=\sum_{i=1}^{n}\lambda_i z(\nu_i)=\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}Z(t)\mathrm{d}t \qquad (6-7)$$

在平稳假设条件下采用线性估计形式,就可算出 $E[Z(V)-Z^*(V)]$和 $E[Z(V)-Z^*(V)]^2$,这样,就可进一步讨论在无偏性和最优性的条件下,如何确定线性估计量的各个权系数 λ_i 的问题(也就是确定函数 f)了。

二、无偏性条件和线性估计方差

1.无偏性条件

采用如同式(6-7)线性估计形式,在满足无偏性和最优性条件下,即可求出权系数 λ_i。无偏性条件要求估计值的期望值与真实值的期望值相等,即 $E[Z^*(V)]=E[Z(V)]$,下面讨论两种情况下无偏性条件的满足问题。

1)当 $E[Z(x)]$为一已知的常数 μ 时的无偏性条件

此时,考察 $Z'(x)=Z(x)-\mu$,由于 $Z(x)$为二阶平稳区域化变量,所以 $Z'(x)$也是二阶平稳区域化变量,且 $E[Z'(x)]=0$。

在估计问题中,只要能对 $Z'(x)$在待估域 V 中的变量值 $Z'^*(V)$进行估计,就可以通过转换 $Z(x)=Z'(x)+\mu$ 对 $Z(x)$在待估域 V 中的变量值 $Z^*(V)$进行估计。此时,我们把式(6-7)转换为如下形式

$$Z'^*(V)=\sum_{i=1}^{n}\lambda_i z'(\nu_i)=\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}Z'(t)\mathrm{d}t \qquad (6-8)$$

现在,我们来讨论无偏性条件的满足问题。

一方面,估计值 $Z'^*(V)$为一随机变量,其期望值为

$$E[Z'^*(V)]=E\left[\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}Z'(t)\mathrm{d}t\right]=\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}E[Z'(t)]\mathrm{d}t=0 \qquad (6-9)$$

另一方面,真实值也是一随机变量,其期望值为

$$E[Z'(V)]=\frac{1}{V}\int_V E[Z'(y)]\mathrm{d}y=0 \qquad (6-10)$$

由式(6-9)和(6-10)可见,$E[Z'^*(V)]=E[Z'(V)]$自然成立,即当 $E[Z(x)]$为一已知的常数 μ 时,无偏性条件自然得到满足。

2)当 $E[Z(x)]$的值未知时的无偏性条件

一方面,估计值 $Z^*(V)$为一随机变量,其期望值为

$$E[Z^*(V)]=E[\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}Z(t)\mathrm{d}t]=\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}E[Z(t)]\mathrm{d}t=\sum_{i=1}^{n}\frac{\lambda_i}{\nu_i}\int_{\nu_i}\mu\mathrm{d}t=(\sum_{i=1}^{n}\lambda_i)\mu \qquad (6-11)$$

另一方面,真实值也是一随机变量,其期望值为

$$E[(Z(V)]=\frac{1}{V}\int_V E[Z(y)]\mathrm{d}y=\mu \qquad (6-12)$$

要使 $E[Z^*(V)]=E[Z(V)]$,必须使

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{6-13}$$

2.无偏性条件满足的前提下线性估计方差的计算公式

在无编性条件满足的前提下,估计方差的计算公式为

$$\begin{aligned}\sigma_E^2 &= E[Z(V) - Z^*(V)]^2 = E\{[Z(V) - \mu] - [Z^*(V) - \mu]\}^2 \\ &= E[Z(V) - \mu]^2 + E[Z^*(V) - \mu]^2 - 2E\{[Z(V) - \mu][Z^*(V) - \mu]\} \\ &= \frac{1}{V^2}\int_V\int_V C(y - y')\mathrm{d}y\mathrm{d}y' + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\lambda_i\lambda_j}{v_i v_j}\int_{v_i}\int_{v_j} C(t - t')\mathrm{d}t\mathrm{d}t' \\ &\quad - 2\sum_{i=1}^{n}\frac{\lambda_i}{Vv_i}\int_V\int_{v_i} C(y - t)\mathrm{d}y\mathrm{d}t\end{aligned} \tag{6-14}$$

令

$$\frac{1}{Vv_i}\int_V\int_{v_i} C(y - t)\mathrm{d}y\mathrm{d}t = \bar{C}(V, v_i)$$

$$\frac{1}{V^2}\int_V\int_V C(y - y')\mathrm{d}y\mathrm{d}y' = \bar{C}(V, V)$$

$$\frac{1}{v_i v_j}\int_{v_i}\int_{v_j} C(t - t')\mathrm{d}t\mathrm{d}t' = \bar{C}(v_i, v_j)$$

则有

$$\sigma_E^2 = \bar{C}(V, V) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\bar{C}(v_i, v_j) - 2\sum_{i=1}^{n}\lambda_i\bar{C}(V, v_i) \tag{6-15}$$

式中,$\bar{C}(V, V)$代表当线段的两个端点分别独立地扫过整个 V 时,所得变差函数的平均值;$\bar{C}(v_i, v_j)$代表当线段的两个端点分别独立地扫过 v_i 和 v_j 时,所得变差函数的平均值;$\bar{C}(V, v_i)$代表当线段的两个端点分别独立地扫过 V 和 v_i 时,所得变差函数的平均值。

用 v 表示所有的 v_i,则有

$$\sum_{i=1}^{n}\lambda_i\bar{C}(V, v_i) = \bar{C}(V, v)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{C}(v_i, v_j) = \bar{C}(v, v)$$

估计方差计算公式(6-15)可简写为

$$\sigma_E^2 = \bar{C}(V, V) + \bar{C}(v, v) - 2\bar{C}(V, v) \tag{6-16}$$

3.关于估计方差的几点说明

(1)当信息域 v 包含在待估域 V 中时,估计方差又称为外延方差,记为 $\sigma_E^2(V, v)$;

(2)不论 v 和 V 是怎样的域,公式(6-16)皆适用;

(3)$\gamma(h) = \frac{1}{2}E[Z(x) - Z(x + h)]^2$ 可看作是用 $Z(x + h)$估计 $Z(x)$的估计方差之半;

(4)σ_E^2 的大小可用来衡量估计量的优劣,σ_E^2 越小,估计值越精确,反之亦然;

(5)估计方差公式是在 $Z^*(V)$为诸 $z(\nu_i)$的线性组合这种很一般情况下推导出来的,推导中必须加上 $\sum_{i=1}^{n}\lambda_i = 1$ 这一前提,这样 $Z^*(V)$就可看成是诸 $z(\nu_i)$以诸 λ_i 为权的加权平均;

(6)σ_E^2 有 4 个影响因素:①V 与 ν 之间的距离;②V 的几何特征;③ν 的几何特征、数量和空间分布;④$\gamma(h)$的特性。

第二节　克立金方法

一、简单克立金方法

当 $E[Z(x)]$为一已知的常数 μ 时,引出简单克立金方法,此时,无偏性条件自然满足,估计问题转化为估计方差最小时,求取式(6-8)中的参数 λ_i,此问题即为求极值问题,我们有

$$\frac{\partial\sigma_E^2}{\partial\lambda_i} = 2\sum_{j=1}^{n}\lambda_j\bar{C}(\nu_i,\nu_j) - 2\bar{C}(V,\nu_i) = 0 \qquad (i = 1,2,\cdots,n) \qquad (6-17)$$

把式(6-17)整理后得

$$\sum_{j=1}^{n}\lambda_j\bar{C}(\nu_i,\nu_j) = \bar{C}(V,\nu_i) \qquad (i = 1,2,\cdots,n) \qquad (6-18)$$

式(6-18)即为简单克立金方程组,其建立过程的实质是最小二乘法的应用。解该方程组,就可得到系数 $\lambda_i(i=1,2,\cdots,n)$,代入式(6-8)就可对 $Z'^*(V)$进行估计,进而利用 $Z(x) = Z'(x)+\mu$,就可对 $Z^*(V)$进行估计,得到的估计值即为简单克立金估计值,记为 $Z_{sk}^*(V)$。

二、普通克立金方法

普通克立金法是一种线性无偏最优估计,采用式(6-7)的线性估计形式,无偏条件为 $\sum_{i=1}^{n}\lambda_i = 1$,"最优"即要求估计方差 σ_E^2 最小,这就涉及在 $\sum_{i=1}^{n}\lambda_i = 1$ 条件满足的前提下,求条件极值的问题。求解条件极值问题要用 Lagrange 乘数法,即构造一个新的函数

$$\begin{aligned} F &= \sigma_E^2 - 2\xi\Big(\sum_{i=1}^{n}\lambda_i - 1\Big) \\ &= \bar{C}(V,V) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\bar{C}(\nu_i,\nu_j) - 2\sum_{i=1}^{n}\lambda_i\bar{C}(V,\nu_i) - 2\xi\Big(\sum_{i=1}^{n}\lambda_i - 1\Big) \end{aligned} \qquad (6-19)$$

式中,ξ 为 Lagrange(拉格朗日)算子。

函数 F 相对于λ_i 和ξ 求导,使导数等于 0,可得到$(n+1)$元方程组

$$\begin{cases} \dfrac{\partial F}{\partial\lambda_i} = 2\sum_{j=1}^{n}\lambda_j\bar{C}(\nu_i,\nu_j) - 2\bar{C}(V,\nu_i) - 2\xi = 0 & (i = 1,2,\cdots,n) \\ \dfrac{\partial F}{\partial\xi} = -2\Big(\sum_{i=1}^{n}\lambda_i - 1\Big) = 0 \end{cases} \qquad (6-20)$$

把上式整理后得到普通克立金方程组

$$
\begin{cases} \sum_{j=1}^{n} \lambda_j \bar{C}(\nu_i, \nu_j) - \xi = \bar{C}(V, \nu_i) \qquad (i = 1,2,\cdots,n) \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{6-21}
$$

普通克立金方程建立过程的实质是最小二乘法的应用。解普通克立金方程组可得到 λ_i $(i=1,2,\cdots,n)$和 ξ，$\lambda_i(i=1,2,\cdots,n)$称为普通克立金系数，把 $\lambda_i(i=1,2,\cdots,n)$代入到线性估计式(6-7)中，即可得到待估域 V 中变量的估计值 $Z^*(V)$，得到的估计值即为普通克立金估计，记为 $Z^*_{ok}(V)$；同时，把 $\lambda_i(i=1,2,\cdots,n)$直接代入到式(6-16)中，又可得到估计方差值，该估计方差称为普通克立金方差，记为 $\sigma^2_{ok}(V)$，即

$$
\sigma^2_{ok}(V) = \bar{C}(V,V) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\bar{C}(\nu_i,\nu_j) - 2\sum_{i=1}^{n}\lambda_i\bar{C}(V,\nu_i) \tag{6-22}
$$

式(6-22)中存在双重求和项，在计算上比较复杂，下面我们推导出普通克立金方差的一个比较简单的表达方式。

在普通克立金方程组[式(6-21)]两端同时乘以 λ_i 并从 $i=1$ 到 n 求和得

$$
\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\bar{C}(\nu_i,\nu_j) = \xi + \sum_{i=1}^{n}\lambda_i\bar{C}(V,\nu_i) \tag{6-23}
$$

把式(6-23)代入式(6-22)中，可得到普通克立金方差的一个简化表达形式

$$
\sigma^2_{ok}(V) = \bar{C}(V,V) - \sum_{i=1}^{n}\lambda_i\bar{C}(V,\nu_i) + \xi \tag{6-24}
$$

三、普通克立金方程组和克立金方差的其它表达形式

1. 使用变差函数时的普通克立金方程组和普通克立金方差

把关系式 $\gamma(h) = C(0) - C(h)$[见式(5-15)]代入到式(6-21)和(6-24)，整理后得到用变差函数表达的普通克立金方程组和普通克立金方差

$$
\begin{cases} \sum_{j=1}^{n} \lambda_j \bar{\gamma}(\nu_i, \nu_j) + \xi = \bar{\gamma}(V, \nu_i) \qquad (i = 1,2,\cdots,n) \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{6-25}
$$

$$
\sigma^2_{ok}(V) = \sum_{i=1}^{n}\lambda_i\bar{\gamma}(V,\nu_i) - \bar{\gamma}(V,V) + \xi \tag{6-26}
$$

2. 当样品承载为点承载时的普通克立金方程组和普通克立金方差

当 v_i 非常小时，可看作点承载或准点承载，普通克立金方程组和普通克立金方差表示为

$$
\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(x_i, x_j) + \xi = \bar{\gamma}(V, x_i) \qquad (i = 1,2,\cdots,n) \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{6-27}
$$

$$
\sigma^2_{ok}(V) = \sum_{i=1}^{n}\lambda_i\bar{\gamma}(V,x_i) - \bar{\gamma}(V,V) + \xi \tag{6-28}
$$

3. 当待估域 V 为点 x 且样品承载为点承载时的普通克立金方程组和普通克立金方差

图6-1所示的问题，在克立金方法体系中，称为块克立金法，即对待估域 V 内的变量值进行估计，其实质是估计出变量在待估域内 V 的平均值。当待估域为点 x 时，且样品承载为点承载时的普通克立金方程组和普通克立金方差有如下形式

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(x_i, x_j) + \xi = \gamma(x, x_i) \qquad (i = 1,2,\cdots,n) \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{6-29}$$

$$\sigma_{ok}^2(x) = \sum_{i=1}^{n} \lambda_i \gamma(x, x_i) + \xi \tag{6-30}$$

4. 普通克立金方程组和普通克立金方差的矩阵表示形式

式(6-29)和(6-30)用矩阵形式表达时，先构筑矩阵

$$[\lambda]=\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \\ \xi \end{bmatrix} \quad [M]=\begin{bmatrix} \gamma(x_1, x) \\ \gamma(x_2, x) \\ \cdots \\ \gamma(x_n, x) \\ 1 \end{bmatrix} \quad [K]=\begin{bmatrix} \gamma(x_1,x_1) & \gamma(x_1,x_2)\cdots\gamma(x_1,x_n) & 1 \\ \gamma(x_2,x_1) & \gamma(x_2,x_2)\cdots\gamma(x_2,x_n) & 1 \\ & \cdots\cdots & \\ \gamma(x_n,x_1) & \gamma(x_n,x_2)\cdots\gamma(x_n,x_n) & 1 \\ 1 & 1 \quad \cdots \quad 1 & 0 \end{bmatrix} \tag{6-31}$$

则普通克立金方程组可写为

$$[K][\lambda] = [M] \tag{6-32}$$

普通克立金方程组的系数可用下式求取

$$[\lambda] = [K]^{-1}[M] \tag{6-33}$$

普通克立金方差表示为

$$\sigma_{ok}^2(x) = [\lambda]^T[M] \tag{6-34}$$

矩阵[K]为数据构形，代表样品点 x_i 之间的空间配置关系，矩阵[M]为估计构形，代表待估点 x 与 x_i($i=1,2,\cdots,n$)之间的空间配置关系。

四、普通克立金计算实例分析

【例6-1】 有一区域化变量 $Z(x)$，其变差函数为

$$\gamma(h) = \begin{cases} 0 & (h = 0) \\ \dfrac{3}{2}\dfrac{h}{200} - \dfrac{1}{2}\dfrac{h^3}{200^3} & (0 < h \leqslant 200\text{m}) \\ 1 & (h > 200\text{m}) \end{cases}$$

设样品点为 x_1, x_2, x_3, x_4(图6-2)，4个样品点相对于待估点 x 对称分布，各样品点与待估点 x 之间的距离为50m，试求用样品点对待估点进行估计的普通克立金系数和普通克立金方差。

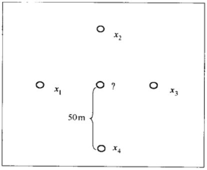


图6-2 例6-1 问题图示

解：建立用矩阵形式表达的普通克立金方程组，需要知道数据构形矩阵[K]和估计构形矩阵[M]的各元素[见式(6-31)]，计算如下：

$$\gamma(x_1,x_2)=\gamma(x_1,x_4)=\gamma(x_2,x_3)=\gamma(x_3,x_4)=\frac{3}{2}\times\frac{70.71}{200}-\frac{1}{2}\frac{70.71^3}{200^3}=0.51$$

$$\gamma(x_1,x_3)=\gamma(x_2,x_4)=\frac{3}{2}\times\frac{100}{200}-\frac{1}{2}\frac{100^3}{200^3}=0.69$$

$$\gamma(x,x_i)=\frac{3}{2}\times\frac{50}{200}-\frac{1}{2}\frac{50^3}{200^3}=0.37\quad(i=1,2,3,4)$$

有了数据构形矩阵[K]和估计构形矩阵[M]后,普通克立金系数和普通克立金方差求取如下:

$$[\lambda]=[K]^{-1}[M]=[0.25,0.25,0.25,0.25,-0.06]$$

$$\sigma_{ok}^2(x)=[\lambda]^T[M]=[0.25,0.25,0.25,0.25,-0.06]\begin{bmatrix}0.37\\0.37\\0.37\\0.37\\1\end{bmatrix}=0.31$$

将求得的系数 $\lambda_i(i=1,2,3,4)$代入式(6-7)即可得普通克立金估计值

$$Z_{ok}^*(x)=\sum_{i=1}^{n}\lambda_i z(x_i)=0.25[z(x_1)+z(x_2)+z(x_3)+z(x_4)]$$

五、普通克立金分析方法的特点

1.普通克立金方程组及方差的特点

(1)解的存在性和惟一性问题:只要 $C(h)$是正定的或 $-\gamma(h)$是条件正定的,则克立金方程组有且只有一组解。

(2)克立金估值是一种无偏的内插估值,在样品点克立金估计值等于实测值,估计方差为0,即

$$Z_{ok}^*(\nu_i)=z(\nu_i)\quad(i=1,2,\cdots n),\quad\sigma_{ok}^2(\nu_i)=0$$

(3)表达式具有通用性:不论数据承载 v_i、v_j 和待估承域 V 的大小和形状如何,不论 $C(h)$ 或 $\gamma(h)$所表征的结构如何,表达式均适用。

(4)普通克立金方程组和方差考虑了以下4方面的因素:① V 的几何特征;② ν_i 与 V 之间的空间配置关系;③数据构形的几何特征;④以变差函数来表现的区域化变量的空间结构特征。

(5)普通克立金估计不依赖于样品数据 $z(\nu_i)$的具体数值,一旦估计构形和变差函数已知,则在开钻之前就可解出普通克立金权系数及相应的方差。这样,就可利用补打新钻孔后克立金方差减少的数量来衡量补打该钻孔的效果,从而确定最优布孔方案。

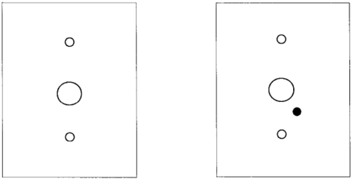
(6)数据构形矩阵[K]只依赖于变差函数和数据构形,在对不同待估域的变量值进行估计时,数据构形是不变的。

(7)普通克立金方程是在线性、无偏和最优条件下推导出来的,从实质上来看,是最小二乘法的应用(见第三章),所以,克立金方程组也可看作是广义的线性回归方程。

2.普通克立金系数的特点

(1)对称性:由变异函数的对称性可以推知,相对于待估点,空间位置对称的两个样品点,其权系数是相同的,如在图 6-3(a)中,两个样品点相对于待估域的位置是对称的,两个样品点的权系数是相同的。

(2)屏蔽效应(screen effect):如图 6-3(b)所示,当有一个样品点靠近待估域时,其作用则屏蔽了离待估域较远样品点的作用,图中下方样品点的作用就受到新加入样品点(实心点)的屏蔽,其克立金权系数减小。



(a)对称性　　(b)屏蔽效应

图 6-3　克立金权系数的特点示意图

(3)克立金权系数可正可负:克立金权系数 λ_i 可为正值,这是最常见的情况,但有时会出现个别 λ_i 为负值的情况。

3.普通克立金法的优缺点

(1)克立金估计比传统方法要精确,至少可以避免系统误差;

(2)不仅给出一个估计值,而且可同时给出估计方差 $\sigma^2_{ok}(x)$ 或 $\sigma^2_{ok}(v)$,而估计方差可作为估计精度的一个衡量指标;

(3)用"统计距离"代替几何距离,能考虑空间点之间的配置以及结构性特征;

(4)可以消除丛集效应:在图 6-4 中,下方有 4 个密集分布的样品点,这 4 个样品点的权系数之和几乎和上方一个样品点的权系数相同;

(5)用结构模型,可以更好地考虑定性信息的影响:在变差函数拟合过程中,人们对变量在各向异性等方面的地质认识,可以指导人们在模型选择等方面的工作,而变差函数对克立金估值有直接影响;

(6)可以直接用一个克立金方程组估计块段上的平均值或某一点上的变量值,非常方便;

(7)系数可正可负,使估值范围可超出实测值范围;

(8)经典统计学考虑的是克立金方法的特例,而克立金方法是经典统计学的发展,当区域化变量是纯随机变量时,克立金估计等同于用经典统计学进行的平均值估计法。

克立金估计方法的缺点是,计算工作量大,然而,随着计算机技术的迅速发展,这一缺点已变得不明显。

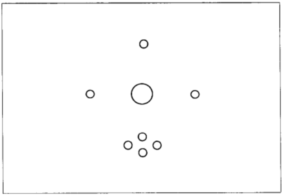


图 6－4　减弱丛集效应示意图

六、普通克立金应用实例分析

塔里木盆地轮南凸起中部超压压力系统的压力分布如图 6－5 所示。

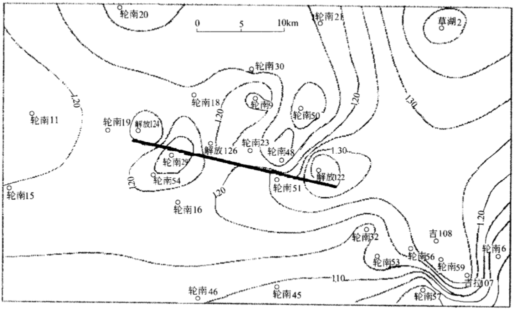


图 6－5　塔里木盆地轮南凸起中部超压压力系统
实测地层压力系数分布图和桑塔木断裂(图中粗实线)所在位置示意图

我们收集了 48 口井的数据，在第五章实例所示的结构分析的基础上，对其中 15 口井的压力值用本章介绍的普通克里金方法进行了估计，并与实测值进行了对比——交叉验证(表 6－1)。由表 6-1 可见，有 10 口井的预测精度大于 90%，有 3 口井的预测精度在 88% ~ 90%之间，只有两口井的预测精度在 86% ~ 88%之间，总的来看，克立金法预测的精度高，能满足压力预测的工程需要。

表 6-1　塔里木盆地轮南凸起中部超压压力系统普通克立金估计值与实测值对照表

井　号	实测值	预测值	绝对误差	预测精度
轮南 39	1.143	1.088	0.055	93.8%
轮南 54	1.070	1.070	0.000	100.0%
轮南 46	1.105	1.085	0.020	98.2%
轮南 32	1.212	1.219	0.070	99.4%
轮南 21	1.161	1.132	0.029	97.5%
轮南 23	1.264	1.251	0.013	99.0%
轮南 19	1.226	1.249	0.023	98.1%
轮南 16	1.205	1.063	0.142	88.2%
轮南 53	1.206	1.112	0.094	92.2%
轮南 51	1.247	1.076	0.171	86.3%
轮南 11	1.261	1.201	0.060	95.2%
轮南 59	1.355	1.216	0.139	89.7%
解放 126	1.206	1.073	0.133	89.0%
吉 108	1.353	1.268	0.065	93.7%
解放 124	1.275	1.107	0.168	86.8%

预测精度稍低的两口井——轮南 51 井和解放 124 井位于桑塔木断裂带附近(图 6-5)。断裂带附近地层的孔渗性得到改善,断裂带又常常是泄压的通道,在断裂带附近压力分布得到一定程度的调整,克立金方法在进行压力预测时,所能考虑的是地层压力区域化变量在其空间域中的整体的结构性状,而无法考虑在断裂带附近地层压力受局部扰动所出现的局部性结构特点,这就导致克立金方法在断裂带附近预测的精度要低一些。

第三节　指示克立金方法

一、指示克立金法的产生

上节讨论的普通克立金方法,是一个十分有效的内插估计方法,对解决变量的空间展布问题很有用。但在实际地质问题中,常常还会遇到另外一类问题,如矿产资源的可采储量计算问题,要解决这类问题,需要知道某一点或某一区域矿石品位大于某一给定值的概率,或者说需要知道 x 点处随机变量 $Z(x)$ 的概率分布(图 6-6),在假设变量不存在空间结构性的条件下,事件 $\{Z(x)\leqslant z_k/(n)\}$ 出现的概率等于事件 $\{z(x_i)\leqslant z_k/(n)\}(i=1,2,\cdots,n)$ 出现的比例。如在图 6-7(a)中,待估点变量 $Z(x)\leqslant 9$ 的概率为 0.6。

但对地质变量而言,绝大多数具有空间相关性,要解决 x 点处随机变量 $Z(x)$ 的概率分布问题,就需要引入新的方法,指示克立金法正是为解决此类问题而发展起来的。

二、指示克立金方法原理

1. 指示转换(指示编码)

现在,对于任意给定的阈值 z_k,引入(定义)指示函数,对每个信息(样品)点 x_i 有

$$I(x_i;z_k)=\begin{cases}0 & if \quad z(x_i)>z_k \\ 1 & if \quad z(x_i)\leqslant z_k\end{cases}\quad (i=1,2,\cdots,n) \qquad (6-35)$$

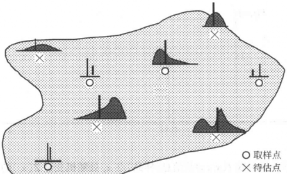


图 6-6　待估点随机变量 $Z(x)$ 的概率图示

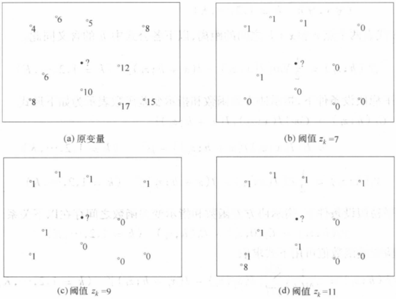


图 6-7　原变量对不同阈值 z_k 的指示转换

原变量[图 6-7(a)]相对于不同的阈值 z_k，有不同的指示转换[图 6-7(b)，6-7(c)，6-7(d)]。

指示转换函数 $I(x_i, z_k)(k=1,2,\cdots,K)$ 有明确的物理含义，它实际上是信息(样品)点 x_i 处随机变量 $Z(x_i)$ 的概率分布函数(图 6-8)，即

$$F(x_i;z_k)=P\{Z(x_i)\leqslant z_k\}$$

$$=I(x_i;z_k)=\begin{cases}0 & 若\quad z(x_i)>z_k\\1 & 若\quad z(x_i)\leqslant z_k\end{cases}\quad(k=1,2,\cdots,K)\qquad(6-36)$$

2. 指示协方差函数和变差函数

指示转换函数同样可看作是一个区域化变量，可与区域化变量相同的形式定义指示转换函数的协方差函数和变差函数，即

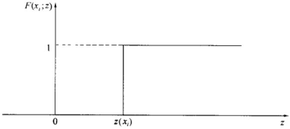


图 6-8　指示转换函数 $I(x_i;z)$是信息(样品)点 x_i 处随机变量 $Z(x_i)$的概率分布函数

$$C_I(h;z_k) = Cov\{I(x;z_k), I(x+h;z_k)\}$$
$$= E\{I(x;z_k) - E[I(x;z_k)]\}\{I(x+h;z_k) - E[I(x+h;z_k)]\}$$
$$(\forall x, \forall h \quad k = 1,2,\cdots,K) \tag{6-37}$$

式中,h 代表两个点 x 和 $x+h$ 之间的距离,以下各公式中 h 的含义同此。

$$\gamma_I(h;z_k) = \frac{1}{2}\mathrm{Var}[I(x;z_k) - I(x+h;z_k)] \quad (k = 1,2,\cdots,K) \tag{6-38}$$

在二阶平稳假设条件下,指示协方差函数和指示变差函数表示为如下形式

$$C_I(h;z_k) = Cov\{I(x;z_k), I(x+h;z_k)\}$$
$$= E[I(x;z_k)I(x+h;z_k)] - \mu^2 \quad (k = 1,2,\cdots,K) \tag{6-39}$$

$$\gamma_I(h;z_k) = \frac{1}{2}E[I(x;z_k) - I(x+h;z_k)]^2 \quad (k = 1,2,\cdots,K) \tag{6-40}$$

在二阶平稳假设条件下,指示协方差函数和指示变差函数之间存在以下关系

$$\gamma_I(h;z_k) = C_I(0;z_k) - C_I(h;z_k) \quad (k = 1,2,\cdots,K) \tag{6-41}$$

指示实验变差函数值可用下式求取

$$\gamma_I^*(h;z_k) = \frac{1}{2N(h)}\sum_{j=1}^{N(h)}[I(x_j;z_k) - I(x_j+h;z_k)]^2 \quad (k = 1,2,\cdots,K) \tag{6-42}$$

式中,$N(h)$为相距为 h 的数据点对的个数。

指示变差函数的拟合与前面已介绍过的区域化变量变差函数拟合的原理和步骤完全一样(见第五章),在此不再赘述。

3.指示克立金方程组

在引入指示转换函数后,假设变量无空间结构性的条件下,x 点处随机变量 $Z(x)$的概率分布函数可由下式估计

$$F\{x;z_k/(n)\} = P\{Z(x) \leqslant z_k/(n)\}$$
$$= \frac{1}{n}\sum_{i=1}^{n}I(x_i;z_k) \quad \forall x \quad (k = 1,2,\cdots,K) \tag{6-43}$$

式中,F 是 z_k 和 n 个已知数 $z(x_i)$　$(i = 1,2,\cdots,n)$的函数。

在变量存在空间结构性的条件下(常见的情况),估计 x 点处随机变量 $Z(x)$的概率分布函数就不这么简单,此时需要引用指示克立金方法。

在考虑变量的空间相关性时，则有

$$F\{x;z_k/(n)\} = P\{Z(x) \leqslant z_k/(n)\} = \sum_{i=1}^{n}\lambda_i(x;z_k)I(x_i;z_k) \quad (k = 1,2,\cdots,K) \quad (6-44)$$

式(6-36)和(6-44)表明，求 $F\{x;z_k/(n)\}$ 可以看作是用已知的 $I(x_i;z_k)$ 来估计未知的 $I(x;z_k)$。

用严格的概率论理论可以证明，$\sum_{i=1}^{n}\lambda_i(x;z_k)I(x_i;z_k)$ 确是已知 n 个数据时 $Z(x)$ 的条件概率分布函数，或是 $I(x;z)$ 的线性估计，即

$$F\{x;z_k/(n)\} = I^*(x;z_k) = \sum_{i=1}^{n}\lambda_i(x;z_k)I(x_i;z_k) \quad (6-45)$$

给定任一点 x 和任一阈值 z_k，可建立指示克立金方程组

$$\begin{cases}\sum_{j=1}^{n}\lambda_j(x;z_k)\gamma_I(x_i - x_j;z_k) + \xi(x;z_k) = \gamma_I(x - x_i;z_k) \\ \sum_{i=1}^{n}\lambda_i(x;z_k) = 1 \quad (i = 1,2,\cdots,n)\end{cases} \quad (6-46)$$

式中，γ_I 代表指示变差函数；$x_i - x_j$ 代表两个样品点之间的距离；$x - x_i$ 代表待估点 x 和样品点 x_i 之间的距离。

解指示克立金方程组(6-46)得到系数 $\lambda_i(x;z_k)(i=1,2,\cdots,n)$ 和 $\xi(x;z_k)$，代入到式(6-45)中，即可求出 $I^*(x;z_k)$。

一般把 $z(x_i)$　$(i=1,2,\cdots,n)$分成 K 个阈值 $z_1,z_2,\cdots,z_K$，故在每一点 x 上，有 K 个指示克立金方程组需求解，可得到 K 个 $I^*(x;z_k)$估计值，这 K 个 $I^*(x;z_k)$估计值分别对应待估点 x 处随机变量 $Z(x)$小于 $z_1,z_2,\cdots,z_K$ 的概率(图 6-9)。

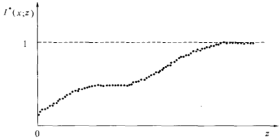


图 6-9　$I^*(x;z_k)$—$z_k(k=1,2,\cdots,K)$关系曲线即为 x 点处随机变量 $Z(x)$的概率分布曲线

对不同的点，可以得到不同的概率分布曲线，由这些不同的概率分布曲线，可读出区域化变量 $Z(x)$在不同点出现小于某一值 z_c 的概率，由此可绘制出等概率图(图 6-10)。也可以根据不同点的概率分布曲线，事先确定一个概率值 P_c，绘制变量等值线图，在变量等值线图上可以读出区域化变量 $Z(x)$以 P_c 为概率的取值(图 6-11)。等概率图和变量等值线图皆反映区域化变量的不确定性。

4. 指示克立金对“软数据”的利用

以上所用到的观测点的数据皆为一确定的实数(见图 6-7)，我们称为“硬数据”，在地质

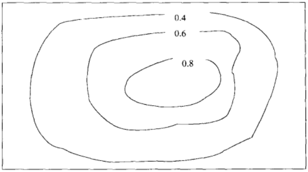


图 6-10　等概率图

图中的值为区域化变量 $Z(x) \leq z_c$ 的概率，z_c 为某一给定的值

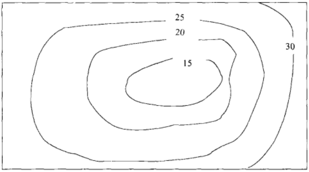


图 6-11　变量取值等值线图

图中的值为区域化变量 $Z(x)$以 P_c 为概率的取值

研究中，还常常碰到另外一类数据，如已知某一点的变量值落在某个区间内 $z(x_i) \in (a, b)$，或服从某一分布 $P\{Z(x_i) \leq z_k\} = F(x_i; z_k) \in (0,1)$，这类数据称为“软数据”。如图 6-11 为图 6-7(a)加入一个新信息(样品)值，该值在 6~8 之间变化(图 6-12 中划底线的值)。

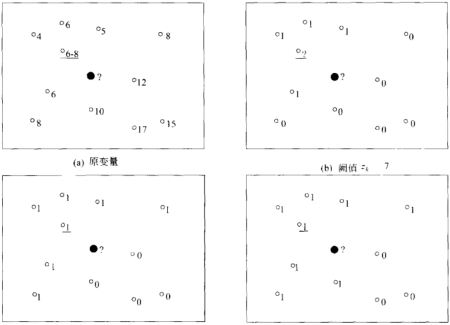
指示克立金可以利用这些软数据中包含的信息，这是其它方法所没有的功能，在指示转换中，这一新的软数据得到了利用，见图 6-12 中的(c)，(d)。

5. 关于指示克立金的几点说明

(1)指示克立金法的目的是给出区域化变量在空间某一点处随机变量的概率分布曲线，从而一方面可对变量的不确定性进行度量，另一方面又可给出变量值大于(或小于)某一给定值的概率，可解决诸如可采储量估计这类问题；在得到一个估计前独立地求出概率分布，以表征变量的不确定性。估计值有多种，但概率分布估计只有一种。

(2)指示克立金法的目的不是得到一个最优估计，而是得到待估点随机变量的概率分布曲线，以便对不确定性进行衡量。在指示克立金中，不确定性以概率分布的形式来表示，与普通克立金法用克立金估计方差表示不确定性是不同的。

(3)指示克立金法是非线性和非参数的估计：尽管指示克立金方程组是线性方程组，但由于指示转换本身是一个非线性转换，所以，指示克立金估计是非线性估计。指示克立金对变量的概率分布模型没有任何要求，变量的分布可以是任一模型，故指示克立金是一种非参数估计。



(a) 原变量　(b) 阈值 $z_k=7$　(c) 阈值 $z_k=9$　(d) 阈值 $z_k=11$

图 6-12　在图 6-7 中加入一个软数据后的指示转换情况

(4)指示克立金法可以利用“软数据”，这是其它方法所不具备的优点。

(5)指示克立金法既可用于连续型变量，也可用于离散型(属性)变量。

三、指示克立金应用实例分析

美国西维吉尼亚中部 Granny Creek 油田在 1924 年～1960 年进行了开发，从下密西西比 Price 组顶部砂岩储层中累计产油 650 万桶，获得了大量开发资料。为了搞清控制 Price 组顶部砂岩储层单井初始产量的地质因素，需要对单井初始产量在平面上的分布进行研究和定量描述。

把 Granny Creek 油田单井初始产量作为区域化变量，图 6-14 为对应不同阈值的指示变差函数，其计算步骤是先按式(6-35)进行指示转换(指示编码)，再按式(6-38)计算实验指示变差函数并进行函数拟合(图 6-13 中的实线)。

按上述步骤，通过指示克立金计算，可得到 Granny Creek 油田 Price 组顶部砂岩储层单井初始产量大于不同阈值的概率等值线图(图 6-14)，图 6-14(a)和(b)分别为单井初始产量大于 10 桶/天和大于 15 桶/天的概率等值线图。

图 6-14 提供了关于 Granny Creek 油田 Price 组顶部砂岩储层单井初始产量分布的综合性信息，反映了单井初始产量在平面上的变化，根据单井初始产量在平面上的变化情况，就可以进一步研究 Price 组顶部砂岩储层单井初始产量的地质控制因素，为邻区该储集层的勘探开发部署提供重要的参考依据。

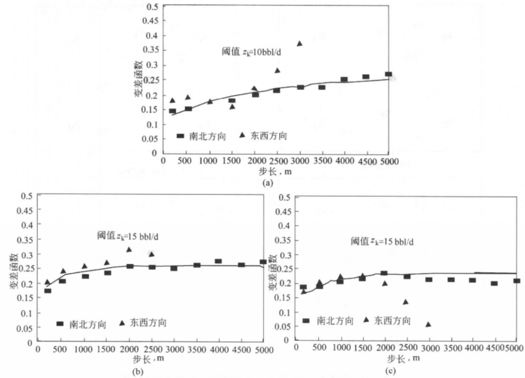
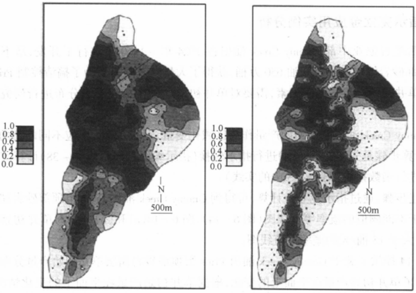


图 6-13 Granny Creek 油田 Price 组顶部砂岩储层单井初始产量区域化变量对应不同阈值的指示变差函数



(a) 初始产量超过去10桶／天的概率等值线　　(b) 初始产量超过去15桶／天的概率等值线

图 6-14 Granny Creek 油田 Price 组顶部砂岩储层单井初始产量大于某一阈值的概率等值线图

◆ 思 考 和 练 习 题 ◆

1.估计的一般形式是什么？为什么要采用线性估计形式？
2.尝试推导普通克立金方程。
3.克立金法有哪些优缺点？
4.为什么说克立金法是一种线性、无偏和最优的估计？
5.克立金方程组有哪些特点？
6.如何理解克立金方差？
7.指示转换的物理含义是什么？
8.指示克立金法有哪些优点？
9.指示克立金和普通克立金有何不同？
10.在图 6－12 中，当新信息(观测)点的取值在 10～13 之间变化时，如何进行指示转换？

第七章　区域化变量的随机模拟

本章学习目的和要求：掌握连续变量、离散变量随机模拟和退火模拟的基本原理，了解其在油气藏参数模拟中的应用。学习本章内容，第五章—第六章的知识是必不可少的，同时也需要了解第二章中联合概率分布的概念。

第一节　区域化变量随机模拟简介

一、模拟及模拟分类

1.模拟

用一种模型来模仿、拟合(地质)现象或再现(地质)过程，称为模拟。

2.模拟分类

模拟有不同的分类方法：按实现的途径和手段，可分为物理模拟和计算机(或数学)模拟；按照模拟的思路不同，可分为确定型模拟和随机型模拟(简称随机模拟)；按照是否有时间因素参入，可分为静态模拟和动态模拟。

1)物理模拟和计算机(或数学)模拟

物理模拟借助于物理模拟装置模仿、拟合(地质)现象或再现(地质)过程，物理模拟具有时间和空间(模型大小)局限性，对复杂地质现象或过程的模拟，只能给出一些参数，或比较适合对(地质)现象或(地质)过程进行机理研究。

计算机模拟借助于计算机模仿、拟合(地质)现象或再现(地质)过程，从理论上讲，可以模拟的时间段和空间大小不受限制，与物理模拟相比，有很大的优越性。

2)确定型模拟和随机型模拟(简称随机模拟)

确定型模拟从事物的本质出发，在获得各变量变化规律及变量之间定量或半定量关系的基础上，构建数学模型，并通过模型的求解，来再现地质过程。

随机型模拟从观察到的数据(现象)出发，在尊重已知数据的情况下，模拟未取样点变量所有可能取值的情况，以分析变量的变化性和不确定性。

3)静态模拟和动态模拟

静态模拟不考虑时间因素，而动态模拟考虑时间因素。

以上不同的分类可组合使用，如确定型动态计算机模型(盆地模拟)、随机型静态计算机模型(油藏描述)和随机型动态计算机模型(油藏模拟)等。

另外，同一分类方案中不同类型的模拟，也可以结合起来使用，如物理模拟和计算机模拟常常结合在一起，确定型模拟和随机型模拟也常常结合在一起。

3.非条件模拟和条件模拟

非条件模拟，只再现变量的结构特征，而条件模拟既能再现变量的结构特征，又能在样品点上使模拟值等于真实值，其最大的好处是能再现区域化变量真实值的离散性和波动性。

二、Monte—Carlo 模拟的基本原理

Monte—Carlo 思想是 Von Neuman 和 Ulam 在研究裂变物质的中子随机扩散时提出的，它利用随机数进行统计试验（相当于取样过程或随机试验），以求得统计特征值作为待解问题的数值解，其基本步骤如下所述。

1.随机数概率分布函数 $F(z)$ 的选择

在生成随机数之前，首先，要确定随机数的概率分布函数 $F(z)$ 及参数，这种分布类型可能是事先知道的或通过实际数据拟合得到的。

2.随机数列的生成

1)均匀随机数的生成

产生在[0,1]区间内均匀分布的随机数 r_i，见图 7 - 1(a)，均匀随机数产生方法有观测法、数学方法等，也可用随机数生成装置来生成。理论上可以生成无穷多个 r_i 值，但实际操作过程中，选一个足够大的 i 值即可，一般选择 $i = 100 \sim 200$，但一般 i 应大于 30，以保证分析具有统计学意义。

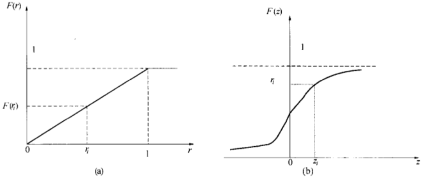


图 7 - 1 [0,1]之间均匀分布随机数 r_i 的生成(a)和服从概率分布 $F(z)$ 的随机数生成(b)

2)服从概率分布函数 $F(z)$ 的随机数的生成

把[0,1]区间均匀分布的随机数 r_i 看作是概率分布函数 $F(z)$ 的概率值，即 $F(z_i) = r_i$，用反变换 $z_i = F^{-1}(r_i)$，即可得到服从分布函数 $F(z)$ 的 z_i，见图 7 - 1(b)。

3.随机数 z_i 的统计特征分析

对随机数 z_i 进行统计分析，以求得统计特征值作为待解问题的数值解。

三、空间变量研究中随机模拟的必要性和意义

克立金估计是一种线性、无偏和最优的估计，具有平滑效应，不能真实地反映变量的变化情况，平滑效应由下列圆滑关系式表示（图 7 - 2）

$$D^2(\nu/G) = D^2_{ok}(\nu/G) + \overline{\sigma^2_{ok}} \qquad (7-1)$$

上式中的 $D^2(\nu/G)$ 是以 ν 为承载的变量在 G 域中的离散方差（定义离散方差为方差的期望值），$D^2_{ok}(\nu/G)$ 为克立金估计值 $Z^*_{ok}(\nu)$ 在 G 域中的离散方差（图 7 - 3）。式(7 - 1)中的 $\overline{\sigma^2_{ok}}$

为各单元上克立金方差之平均值，其表达式为

$$\overline{\sigma_{ok}^2} = \frac{1}{n}\sum_{i=1}^{n} \sigma_{ok}^2(\nu_i) = \frac{1}{n}\sum_{i=1}^{n} E\left[z(\nu_i) - z_{ok}^*(\nu_i)\right]^2 \tag{7-2}$$

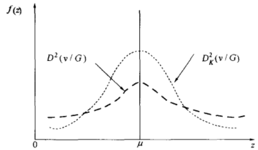


图 7-2　克立金估计平滑效应示意图

正是因为克立金方法的平滑效应，使得克立金估计把变量的变化性以低通滤波的形式滤掉，从而掩盖或不能很好地再现变量的变化性，而对这种变化性的认识有时是非常重要的，如在油藏范围内，一个高孔渗通道对开发来说可能很重要，由背景区过渡到高孔渗通道区，变量值发生了较大的变化，克立金方法不能有效地反映这种变化，而随机模拟给出的模拟图形，能反映变量的这种变化性。

另一方面，由随机模拟给出的反映变量不确定性的随机图形，输入到其它模型中（如黑油模型中），获得研究变量或称输出变量（如见水时间）的概率分布曲线，为风险决策提供科学依据（图 7-4）。

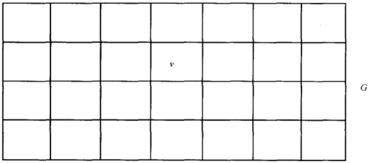


图 7-3　关于离散方差的示意图

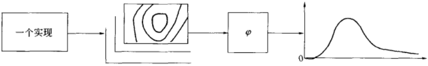


图 7-4　随机模拟过程和应用示意图

四、随机建模技术在油藏描述中的应用情况

随机建模第一次报道是在 1984 年，如今随机建模技术和露头分析、高分辨率地震结合共同构成了储层研究中的新技术。使用统计概念和随机性概念来建立地质模型，把油藏工程研究提高到了一个新的水平，建立储层随机地质模型的过程简称为储层的随机建模。

早期的储层研究主要利用定性或半定量的确定型地质模型，称为储层描述（Reservoir description）。利用充分的地质资料结合现代新技术对储层分布的复杂性（非有序性、各向异性、

非连续性和非均质性等)进行定量化评价,称为储层表征(Reservoir Characterization)。

由于获取的地质资料和信息、数据具有分散性、稀疏性、跳跃性、间接性和随机性,利用这些资料来建立模型时,其结果必然包含有不确定性和多解性,这就要求在储层表征中引入随机建模技术。

随机建模用于研究储层静态行为,具体包括下列几个方面:(1)岩相的划分和空间分布;(2)沉积相带划分和空间分布;(3)孔、渗、饱的空间分布;(4)砂体中悬挂着的泥岩的空间分布和尺寸;(5)储层中遮挡带的空间分布;(6)不同砂体之间的连通性;(7)储层的空间位置和形状;(8)断层和裂缝的位置、方向及其空间分布;(9)储层可采油气的空间分布;(10)储层的产量预报;(11)油田开发中的风险评价;(12)提供油田开发方案。

第二节　连续变量的随机模拟

一、估计值加模拟误差模拟方法

1.基本原理

设 $Z_{ok}^*(x)$为区域化变量 $Z(x)$在 x 点处的克立金估计值,通过简单变换可得

$$Z(x) = Z_{ok}^*(x) + [Z(x) - Z_{ok}^*(x)] = Z_{ok}^*(x) + R(x) \qquad (7-3)$$

根据转换带法等,可以得到区域化变量 $Z(x)$的非条件模拟 $Z_s(x)$,由此可读出在样品点 x_α 处的非条件模拟值 $Z_s(x_\alpha)(\alpha=1,2,\cdots,n)$,根据 $Z_s(x_\alpha)(\alpha=1,2,\cdots,n)$,可用克立金法得到估计值 $Z_{sok}^*(x)$。

由于 $Z_s(x)$与 $Z(x)$同构(即两者具有相同的数学期望、相同的分布直方图和相同的$C(h)$或 $\gamma(h)$),且估计构形相同,故知其克立金权系数 λ_α 也一样,即

$$Z_{ok}^*(x) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha),\ Z_{sok}^*(x) = \sum_{\alpha=1}^{n} \lambda_\alpha Z_s(x_\alpha)$$

故 $Z_{sok}^*(x)$与 $Z_{ok}^*(x)$是同构的,显然,$Z_s(x)-Z_{sok}^*(x)$与 $Z(x)-Z_{ok}^*(x)$也是同构的。

同时,由于 $Z(x)-Z_{ok}^*(x)$与 $Z_{ok}^*(x)$独立,故非条件模拟 $Z_s(x)-Z_{sok}^*(x)$也与 $Z_{sok}^*(x)$独立。

用一个与未知克立金误差 $Z(x)-Z_{ok}^*(x)$同构且独立的非条件模拟的克立金误差 $Z_s(x)-Z_{sok}^*(x)$来代替未知的克立金误差 $Z(x)-Z_{ok}^*(x)$,就可得到条件模拟公式

$$\begin{aligned} Z_{sc}(x) &= Z_{ok}^*(x) + [Z_s(x) - Z_{sok}^*(x)] \\ &= Z_s(x) + \sum_{\alpha=1}^{n} \lambda_\alpha [Z(x_\alpha) - Z_s(x_\alpha)] \end{aligned} \qquad (7-4)$$

可以证明式(7-4)中的 $Z_{sc}(x)$具有如下性质 (具体证明见王仁铎等,1988):

(1)$Z_{sc}(x)$与 $Z(x)$有相同的期望值;

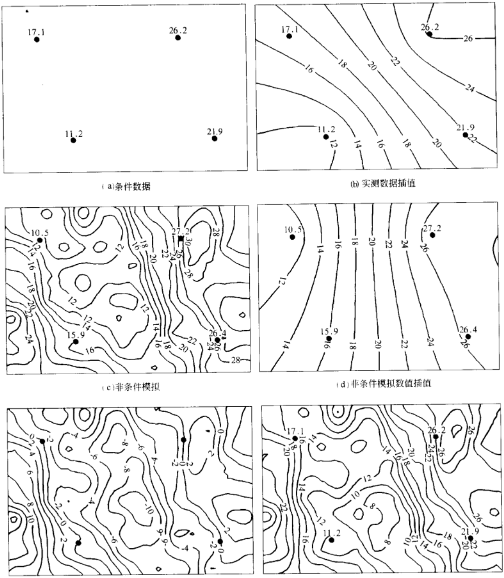
(2)$Z_{sc}(x)$与 $Z(x)$有相同的变差函数;

(3)在实测点上 $Z_{sc}(x_\alpha)=Z(x_\alpha)$;

(4)$E[Z(x)-Z_{sc}(x)]^2 = 2\sigma_{ok}^2$。

根据以上性质,利用式(7-4)就可对区域化变量 $Z(x)$进行条件模拟。

图 7-5 为估计值加模拟误差模拟方法的一个实例。



(a)条件数据　　(b) 实测数据插值

(c)非条件模拟　　(d)非条件模拟数值插值

(e)模拟误差:(c)±(d)　　(f)条件模拟:(b)+(e)

图 7-5　估计值加模拟误差模拟方法实例

图 7-5(a)为实测数据，作为模拟的条件，故又称为条件数据；图 7-5(b)为实测数据的克立金插值，对应式(7-4)中的 $Z^*_{ok}(x)$；图 7-5(c)是根据条件数据[图 7-5(a)]实现的非条件模拟，对应式(7-4)中的 $Z_s(x)$；图 7-5(d)是非条件模拟数据的克立金插值，对应式(7-4)中的 $Z^*_{sok}(x)$；图 7-5(e)为模拟误差，对应式(7-4)中的$[Z_s(x)-Z^*_{sok}(x)]$；图 7-5(f)为条件模拟结果，对应式(7-4)中的 $Z^*_{ok}(x)+[Z_s(x)-Z^*_{sok}(x)]$。

二、序贯模拟

1.基本思路

把样品点的数据或信息称为条件数据，如图 7－6 中的数据即为条件数据。

在模拟时，对研究区首先进行网格化处理，网格节点为待模拟的点(图 7－6)。随机选取一网格节点，对该点的变量值进行模拟，得到一模拟值，然后，在模拟以后各点的变量值时，把先前的模拟值作为条件数据，具体步骤如下：

(1)对研究区进行网格化处理；

(2)随机地选取一个网格节点作为模拟点；

(3)估计该模拟点处随机变量的条件概率分布，即以条件数据作为信息而估计得到的模拟点处随机变量的概率分布；

(4)从该条件概率分布中随机拾取一个值，作为该点随机变量的模拟值；

(5)把(4)中获得的模拟值加入到条件数据中，成为以后模拟中的条件数据；

(6)重复以上步骤直至得到所有网格节点上的模拟值。

完成以上步骤，得到一幅随机图像，即完成了一次随机模拟。

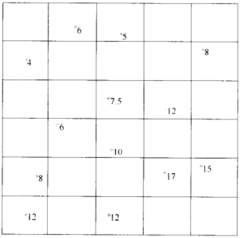


图 7－6　随机模拟网格化图示

重复以上 6 个步骤，可以得到另一幅随机图像，如此进行下去，可得到无穷多幅随机图像。从理论上来讲，这些随机图像出现的可能性是完全相等的。

在实际模拟中，一般获得若干幅随机图像，通过对这些随机图像进行分析、归类，从中筛选出具有代表性的几幅随机图像作为模拟的结果，用来对区域化变量的不确定性进行表征和研究。

2.序贯高斯模拟

1)序贯高斯模拟过程

所谓序贯高斯模拟，是假设区域化变量服从高斯(正态)分布，从而在模拟点处随机变量的条件概率分布为一正态分布，其期望值为该点的克立金估计值 $Z_{ok}^{*}(x)$，其方差为该点的克立金估计方差 $\sigma_{ok}^{2}(x)$，序贯高斯模拟的具体步骤如下：

(1)对研究区进行网格化处理；

(2)随机地选取一个网格节点作为模拟点；

(3)对模拟点进行克立金估计，得到 $Z_{ok}^{*}(x)$ 和 $\sigma_{ok}^{2}(x)$，该点处随机变量为一正态分布，其期望值为 $Z_{ok}^{*}(x)$，其方差为 $\sigma_{ok}^{2}(x)$，见图 7－7(c)；

(4)从该点正态分布变量中随机拾取一个值，作为该点随机变量的模拟值[图 7－7(d)]；

(5)把(4)中获得的模拟值加入到条件数据中，成为以后模拟中的条件数据；

(6)重复以上步骤直至得到所有网格节点上的模拟值。

完成以上步骤，得到一幅随机图像，即完成了一次随机模拟。

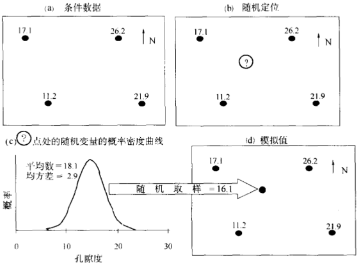


图 7－7 序贯高斯模拟

重复以上 6 个步骤，可以得到另一幅随机图像，如此进行下去，可得到无穷多幅随机图像。

2)序贯高斯模拟的性质

(1)再造一元变量的分布函数 *cdf*（常常以直方图来表示）；

(2)在样品点模拟值等于实测值；

(3)再造协方差函数 $C(h)$或变差函数 $\gamma(h)$。

3.序贯指示模拟

1)方法的原理

两个随机变量 $Z(x)$和 $Z(x+h)$之间的空间相关性可由二维联合分布函数表征

$$P\{Z(x)\leqslant z, Z(x+h)\leqslant z'\}=F(h;z,z') \tag{7-5}$$

在区域化变量 $Z(x)$的协方差函数 $C(h)=E[Z(x)Z(x+h)]-\mu^2$[见式(5－13)]中，$E[Z(x)Z(x+h)]$项可以表示为

$$E[Z(x)Z(x+h)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zz'd^2F(h;z,z') \tag{7-6}$$

故协方差函数 $C(h)$不能完全表征二维联合分布函数 $F(h;z,z')$。

序贯高斯模拟只能再现 $Z(x)$的协方差函数 $C(h)$或变差函数 $\gamma(h)$，不一定能够保证识别出连通性函数 $F(h;z,z')$，与高斯模型有关的模拟具有以下弱点：

(1)不能完全表征二维分布函数 $F(h;z,z')$；

(2)缺乏灵活性，不能再造指示协方差函数；

(3)假设变量服从高斯分布，统计熵最大，即平均信息量最大，无序性最强，故不能识别极值连通性。

因为高斯模型的局限性，在一些应用中，人们提出改变后的两条模拟性质：

(1)再造二维联合分布函数 $F(h;z,z')$；

(2)模拟值可再现协方差函数

由此，产生了序贯指示模拟(SIS)，该方法既可适用于连续变量，又可适用于离散变量，同时，适用于有软/模糊数据的情况。

2)实施步骤

(1)所有已知数据(包括软数据)的指示值编码，见式(6－35)；

(2)K 个指示协方差/变差函数的计算，见式(6－39)和(6－40)；

(3)在模拟网格中的每个节点 x_l，解 K 个指示克立金方程组，见式(6－46)，求出 x_l 点处随机变量 $Z(x_l)$ 以 n 个样品点数据为条件的后验概率分布函数

$$I^*(x_l;z_k) = P\{Z(x_l) \leqslant z_k/(n)\} = \sum_{i=1}^{n}\lambda_i(x_l;z_k)I(x_i;z_k)(k = 1,2,\cdots,K) \tag{7-7}$$

式中，$\lambda_i(x_l;z_k)(i=1,2,\cdots,n)$ 代表对模拟点 x_l 进行估计的指示克立金系数，$I(x_i;z_k)$ 代表样品点 x_i 处变量值的指示编码值；

(4)从后验概率分布函数中随机抽取一模拟值 $z(x_l)$，并把该模拟值加到条件数据中；

(5)返回到(3)重复运算，直到抽过所有的 N 个模拟节点，通过求解 $K \times N$ 次指示克立金方程组，可得到一个实现 $z(x_l)(l=1,2,\cdots,N)$，一个实现给出一幅随机图像。

要获得 M 个实现，需要进行 $K \times N \times M$ 次指示克立金方程组求解。M 个实现给出 M 幅不同的随机图像，这 M 幅不同的随机图像出现的概率是相等的，M 幅随机图像之间的差异反映了变量的空间联合不确定性。

说明：由于指示克立金的性质，改变后的两条性质可以满足，推理过程如下。

考虑指示克立金估计

$$I^*(x_l;z_k) = \sum_{i=1}^{n}\lambda_i(x_l;z_k)I(x_i;z_k) \tag{7-8}$$

对应的简单指示克立金方程组(以指示协方差函数表示)写成

$$\begin{cases}\sum_{j=1}^{n}\lambda_j(x_l;z_k)C_I(x_i - x_j) = C_I(x_l - x_i)(i = 1,2,\cdots,n)\\ \sum_{i=1}^{n}\lambda_i(x_l;z_k) = 1\end{cases} \tag{7-9}$$

指示克立金估计的期望值

$$E\{I^*(x_l;z_k)\} = E\Big\{\sum_{i=1}^{n}\lambda_i(x_l;z_k)I(x_i;z_k)\Big\} = P_k(x_l) = F(x_l;z_k) \tag{7-10}$$

满足改变后的第一条性质，直方图将被模拟值重现。

现在来看指示估计 $I^*(x_l;z_k)$ 与任何一个指示值的协方差函数

$$E\{[I^*(x_l;z_k) - P_k(x_l)][I(x_i;z_k) - P_k(x_i)]\}$$

$$= E\Big\{\Big[\sum_{j=1}^{n}\lambda_j(x_l;z_k)I(x_j;z_k) - E\Big(\sum_{j=1}^{n}\lambda_j(x_l;z_k)I(x_j;z_k)\Big)\Big][I(x_i;z_k) - P_k(x_i)]\Big\}$$

$$= \sum_{j=1}^{n}\lambda_j(x_j;z_k)E\{[I(x_j;z_k) - P_k(x_j)][I(x_i;z_k) - P_k(x_i)]\}$$

$$= \sum_{j=1}^{n}\lambda_j(x_l;z_k)C_I(x_i - x_j) = C_I(x_l - x_i) \tag{7-11}$$

故指示克立金估计 $I^{*}(x_l;z_k)$满足所有的指示协方差函数 $C_I(x_l-x_i)$，改变后的第二条性质满足。

由以上性质可知，序贯指示模拟可以再现区域化变量的连通性，对于模拟油气藏中的高(或低)孔渗带具有重要作用。

三、区域化变量随机模拟的特点

1.与 Monte—carlo 模拟对比

区域化变量的随机模拟借助于 Monte—carlo 模拟的基本思想，不仅能再现区域化变量的概率分布，而且能再现区域化变量的空间结构性特征，而 Monte—carlo 模拟只能再现变量的概率分布。

2.与克立金估计对比

(1)重现变化性，由于通过模拟可求出与 $Z(x)$一样的协方差函数或变差函数，从而可重现真实值的离散方差；

(2)在对不确定性的估计上，普通克立金估计给出一个估计值和相应的估计方差，一次给出一个未知量的不确定性估计，指示克立金用后验条件分布来估计不确定性，而随机模拟给出联合空间分布的估计，能够从整体上对未取样点提供一种不确定性的度量，而不是局部地、一个位置接着一个位置地反映这种不确定性。

空间联合分布能给出一串值 $Z(x_l)(l=1,2,\cdots,N)$大于某个阈值(门槛值)z_c 的概率，如表示一个断层或流体流动的优势通道

$$Z[x_l, l=1,2,\cdots,N/(n)] = P\{z(x_l) > z_c, l=1,2,\cdots,N/(n)\} \qquad (7-12)$$

因此，随机模拟在油气藏开发地质研究中有着广泛的应用。

3.随机模拟更好体现了区域化变量的基本思想

随机模拟更好地体现了区域化变量的基本思想，即空间上任何一点都对应一个随机变量，随机模拟能够再现这些随机变量的联合分布。

4.模拟值不等于真实值

模拟不能制造信息，模拟值不等于真实值，但模拟能根据已知的信息给出变量的变化性和不确定性方面的综合评价。

第三节　离散变量的随机模拟——面向对象的模拟

一、离散变量随机模拟的原理

前面，我们介绍了连续变量随机模拟的原理，但在实际研究中，可能会遇到离散变量随机模拟的问题，如沉积微相类型、岩性分布等问题，这类变量也称为类型变量或属性变量，这就需要有针对这类变量的模拟方法。

对离散变量(类型变量)进行模拟时，我们把类型变量中的每个类型作为一个整体，按一定的规则随机地模拟每个类型的空间位置和测度(长度、面积或体积)，所以，这类模拟方法又称为面向对象的模拟方法。

通过随机地模拟每个类型的空间分布和测度(长度、面积或体积)，最终使各类型的分布和测度达到事先确定的目标，即可得到模拟图像，该图像即为离散变量的一个随机图像，重复模

拟过程，就可得到无穷多个随机模拟图像。

离散变量的随机模拟以 Boolean 过程为基础，Boolean 过程是由几何体在空间上服从于某一概率分布而产生的，示性点过程是 Boolean 过程的扩展和推广，在示性点过程模拟中，物体中心的空间分布构成一个点过程，这一方法不适用于连续变量，不能以显式再现协方差或变差函数，也不能以显式服从已知数据。

同时，我们注意到，第六章中的指示转换采用以下形式也可对离散变量进行研究

$$I(x_i;\text{类型}-k)=\begin{cases}1 & \text{若 } x_i \text{ 点处变量属类型}-k \quad (i=1,2,\cdots,n)\\ 0 & \text{若 } x_i \text{ 点处变量类型不属类型}-k \quad (k=1,2,\cdots,K)\end{cases} \tag{7-13}$$

式中，n 为样品点个数；K 为离散变量类型个数。

经过式(7-13)的指示值编码后，离散变量在取样点就转化成 0 和 1，用前面的序贯指示模拟方法就可实现对该离散变量的模拟，得到的模拟图像即为离散变量的 K 个类型在空间上的分布情况。

在具体应用时，离散变量的随机模拟又可以和连续变量的随机模拟联合使用，即先对类型变量(如岩性类型)进行模拟，然后，在各类型内部用连续变量随机模拟方法实现某一参数(如孔隙度)的随机模拟，这就是所谓的两步模拟方法，该方法有助于提高参数模拟的精度。

二、Boolean 模拟基本步骤——以井间砂体分布模拟为例

现以井间砂体的分布模拟为例，来说明 Boolean 模拟的基本步骤(图 7-8)。

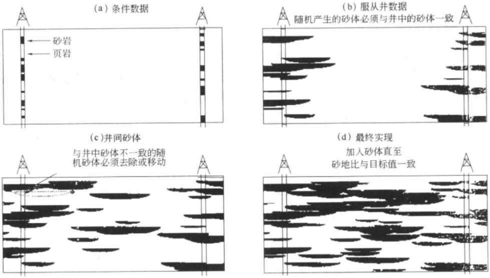


图 7-8　布尔模拟示意图

(1)把两口井钻遇的砂体分布信息作为条件数据，见图 7-8(a)；

(2)假定砂体的形态为向下凸的长轴半透镜体，随机确定其位置和尺寸并置于图中，若模拟出来的信息与条件数据(两口井的信息)不矛盾，则接受该模拟，见图 7-8(b)；

(3)继续前述步骤，若模拟出来的信息与条件数据(两口井的信息)出现矛盾，则删除该模拟，见图 7-8(c)；

(4)重复以上步骤至一事先确定的目标[如砂地比达到某一确定值，图 7-8(d)]。

第四节　退火模拟——二次建模

一、基本原理

“退火”一词源自于热动力学中的一个概念，液(流)体凝固和结晶或金属冷却都是退火过程。高温条件下，分子可自由移动，随着温度逐渐降低，分子组合在一起形成晶体，代表系统的最低能量状态。温度降低，负熵增加，使系统更有序。Boltzman 概率分布表明在一定温度下处于热平衡态的系统的不同热状态 E 出现的概率

$$P(E) \sim e^{-\frac{E}{K_b T}} \tag{7-14}$$

式中，T 为系统的温度；E 为系统的热状态(或称能量状态)；K_b 为 Boltzman 常数。

根据式(7-14)可知，当 E 一定时，T 越高，$P(E)$越大，反之亦然；T 一定时，E 越大，$P(E)$越小，反之亦然。

式(7-14)表明，甚至在低温条件下，也有一个系统处于高能量状态的概率，只不过概率趋于小值。

退火模拟是一种联合优化方法，其实质是寻求具有大量元素的系统的最优组合顺序。

退火算法中的基本问题包括目标函数、目标函数调整方法、干预机制和一些用于降低温度或减弱控制参数的一些经验方法。

二、基本过程

(1)生成一初始图形(这可以是前面模拟的结果或图形)；

(2)建立初始控制参数 C_0 以及循环过程中降低参数 C 的“时间表”(模拟退火进程 Annealing schedule)，初始控制参数的大小可以等于初始目标函数的大小；

(3)干扰图形，如调换任意两个节点处的数值；

(4)计算一个新的目标函数 O_{new}，如目标函数可以是图形变差函数和目标变差函数之差的平方和

$$O = \sum_h \left[\frac{\gamma_i(h) - \gamma_m(h)}{\gamma_m(h)} \right]^2 \tag{7-15}$$

式中，$\gamma_i(h)$为图形的变差函数；$\gamma_m(h)$为目标变差函数。

(5)建立接受概率分布

$$P\{接受\} = \begin{cases} 1 & 若\ O_{new} \leqslant O_{old} \\ e^{\frac{O_{old} - O_{new}}{C}} & 其他 \end{cases} \tag{7-16}$$

(6)根据此概率分布函数做出决定是否接受干扰，若干扰被接受，调整图像和目标函数使 $O_{old} = O_{new}$；

(7)回到 3，循环运算直到目标函数足够小，或在很多循环计算中目标函数无明显变化。

控制参数 C 不能降低得太快，否则图形就不会收敛；若控制参数 C 或温度降低得太慢，则运算时间太长。

图 7-9 为退火模拟的一个实例，图 7-9(a)为原始图形，图中的网格大小为 20m × 3m，黑

色代表砂体,原始图形的实际砂地比为 69.7%,实际砂体平均长度为 29.78m,平均厚度为 4.53m。要求用退火模拟方法调整原始图形,其目标是最终使图形的砂地比为 70%,砂体平均长度为 60m,平均厚度为 10m;图 7-9(b)为第 1 次迭代,随机地选取两个网格(2,17)和(25,25),交换两个网格中的"内容",获得新的图形,计算图形的砂地比为 69.7%,砂体平均长度为 30m,平均厚度为 4.57m,该图形向目标靠近,结果改善,保存图形并进入下一步迭代;图 7-9(c)为第 2 次迭代,随机地选取两个网格(2,13)和(23,5),交换两个网格中的"内容",获得新的图形,计算图形的砂地比为 69.7%,砂体平均长度为 29.34m,平均厚度为 4.45m,该图形远离目标,结果变差,取消该迭代并回到上一步迭代;图 7-9(d)和图 7-9(e)分别为第 463 和 464 次迭代,迭代一直进行下去,直到获得的图形参数与目标参数一致,见图 7-9(f),即完成一次退火模拟。该过程可无限重复,得到无穷多个退火模拟图形,这些图形是等可能出现的。

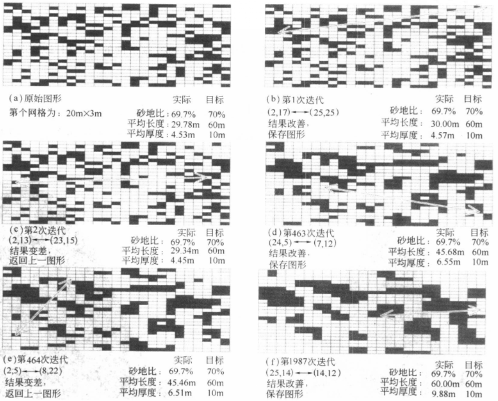


图 7-9 退火模拟实例

第五节 随机模拟应用实例分析

一、地质概况

我国东部某油田下第三系沙河街组的沙三段是油田的主要含油层系,整个沉积环境为扇三角洲的沉积体系,沉积物近源快速堆积,分选差,储层岩性为含砾砂岩、粗砂岩、细砂岩等岩性,横向变化快。储层受沉积条件、构造和成岩作用的影响,非均质性很强,孔隙度分布范围为

10% ~ 29%，渗透率分布范围(30 ~ 290) × $10^{-3}\mu m^2$，储层的非均质性严重影响了开发效果。因此准确预测储层参数的井间变化，揭示储层的宏观非均质性，对于分析油田剩余油分布规律、提高油田开发效果具有十分重要的地质意义，随机建模正是实现上述目的的有效方法。

在储层随机建模中，首先要取得有关储层参数在一定范围内的变化和可靠数据，目前这些数据主要包括钻井资料、测井资料、岩心分析资料以及生产动态资料等。油田现有各类钻井40余口，应用测井资料，通过二次数字处理，可以得到井点处储层的参数。把井点处储层参数作为控制点，通过实验变差函数的计算，选用相应的理论模型进行变差函数的拟合，然后优选模拟算法对井间储层参数进行计算，得到储层参数的三维数据体，从而建立起储层参数的三维模型。下面以该油田储层孔隙度的模拟说明随机建模的过程及实现。

二、建模过程

1. 数据准备

主要包括井位坐标、地质分层、测井解释的孔隙度和作为控制条件的沉积相、地质分层、断层等数据。

2. 孔隙度的统计分析

图7-10和图7-11分别为工区孔隙度概率分布图和累积概率分布图。从图中可以看出，孔隙度的分布范围是0% ~ 30%，优势分布范围是10% ~ 20%，并且孔隙度的分布基本符合正态分布。

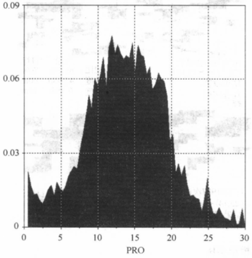


图7-10　孔隙度概率分布图

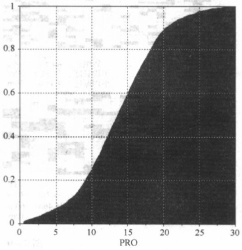


图7-11　孔隙度累积概率分布

3. 实验变差函数的计算及理论变差函数模型的拟合

在数据统计分析的基础上，可以计算任意两口井之间孔隙度的实验变差函数，然后应用理论变差函数模型进行拟合，得到变差函数模型。图7-12为工区孔隙度垂直方向实验变差函数的计算结果和应用球状模型进行拟合得到的变差函数模型。从模型中可以看出，变差函数的变程为23m，基台值19，块金常数为0，于是该模型的表达式为

$$\gamma(h) = \begin{cases} 0 & (h = 0) \\ 19(\frac{3}{2}\frac{h}{23} - \frac{1}{2}\frac{h^3}{23^3}) & (0 < h \leqslant 23) \\ 19 & (h > 23) \end{cases}$$

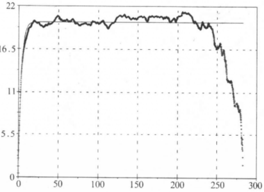


图 7－12　垂直方向孔隙度实验变差函数及其模拟结果

程序还将计算各个方向的变差函数，以便对不同方向孔隙度变化的变差函数进行结构分析，从而得到变差函数的套合结构，为孔隙度的井间预测提供基础。

4.优选相应的模拟方法进行计算，实现孔隙度的井间预测

由于测井资料二次数值处理的精度较高，认为井点处孔隙度值是可靠的，要求模拟结果必须忠实于井点数据，因此选用条件模拟的方法。为了尽量利用已知信息和提高模拟精度，应用序贯模拟的方法进行迭代计算。在实际模拟过程中，针对孔隙度的变化情况，在各种方法实验的基础上，应用序贯高斯模拟的方法进行孔隙度的计算效果较好，网格的几何尺寸为 $50 \times 50 \times 0.25$m，模拟的结果为孔隙度三维数据体，可以沿任意方向进行切片，以研究孔隙度的变化规律。图 7－13 和图 7－14 分别为孔隙度随机模拟的剖面和平面图，从图中可以直观地观察到孔隙度在空间的变化。

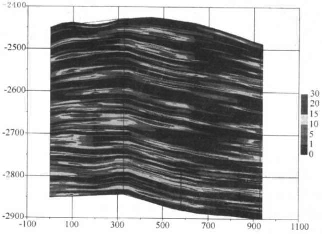


图 7－13　孔隙度随机模拟垂直剖面图

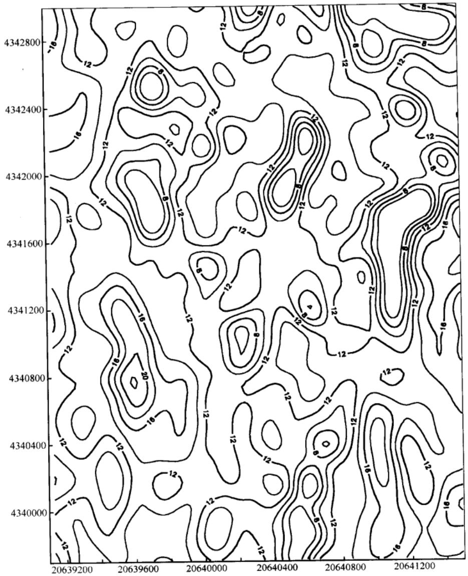


图 7－14　孔隙度随机模拟水平切片等值线图

◆ 思 考 和 练 习 题 ◆

1. 随机模拟有何作用？
2. 随机模拟和普通克立金、指示克立金有何不同？
3. 简述序贯模拟的基本步骤。
4. 序贯高斯模拟与序贯指示模拟的区别是什么？
5. 随机模拟有哪些特点？
6. 简述退火模拟的基本原理。

第八章　非线性科学和分形几何及其应用

本章学习目的和要求：了解非线性科学的基本思维观，掌握分形理论的基本概念和方法原理，针对不同类型分形体，掌握其分形维数的各种计算方法。了解分形理论在油气勘探开发中的应用范围和解决问题的能力，掌握统计分形在实际问题中的多种表现形式。

第一节　非线性科学的认识论

从老三论SCI(贝塔朗菲的一般系统论、维纳的控制论和申农的信息论)到新三论DSC(普里高津的耗散结构论、哈肯的协同论和托姆的突变论)是20世纪科学发展史上的重大革命。耗散结构理论、协同论、突变理论、混沌动力学、分形理论几乎都是创立于50年代，它们无一不是跨学科的，也无一不是非线性的，它们的共同目标是探索大自然中的复杂性，它们从不同的角度揭示出复杂现象中的规律性。这些非线性理论在20世纪是继量子力学和相对论两次重大科学发现后的第三次科学革命。非线性科学正在改变人们对世界的看法，形成了一种新的自然观，促进一大类新兴学科的孕育和发展，并从根本上影响着现代科学的逻辑体系。

天体演化、地球变迁、生命起源，都涉及由大量粒子组成的复杂系统的演变规律。无论是宇宙还是生命，物质世界都经历着从无组织的混乱状态向不同程度有组织状态的演变，实现着从“无序”到“有序”、从“简单”到“复杂”的各种过程，如何从总体上认识自然界发生的这些复杂现象，找出其基本规律，就构成了正在兴起的跨学科研究领域——非线性科学的基本内容。

地球系统是地球科学的研究对象，它是一个复杂的巨系统，非平衡性、非线性、多尺度性、突变性、自组织性、自相似性、有序性和随机性等是其最本质的属性，因而非线性理论及有关方法是研究地球科学和地质学的最有效的理论和方法之一。

一、系统及其分类

1. 系统、线性与非线性

1)系统

系统是由相互作用和相互依赖的若干组成部分结合而成的具有特定功能的整体(钱学森，1978)。系统具有多种属性，由此决定了系统的多种类型。系统有严格的等级和层次之分。美籍匈牙利学者欧文·拉兹洛是继贝塔朗菲之后在美国推动“系统运动”的著名系统哲学家，他用有序整体(α)、自稳定(β)、自组织(γ)和等级性(δ)4个基本构成因素，建立了“基本系统”的普遍模型，即

$$R = f(\alpha, \beta, \gamma, \delta) \tag{8-1}$$

2)线性系统

线性系统具有叠加性，即当z为系统在输入为x时的输出，则有

$$ax + b = az + b \tag{8-2}$$

3)非线性系统

非线性系统具有相干性，即当z为系统在输入为x时的输出，则有

$$ax + b \neq az + b \quad (8-3)$$

在非线性的世界里,随机性和复杂性是其主要特征,但同时,在这些极为复杂的现象背后,存在着某种规律性。严格地说,自然界的系统皆为非线性系统,但有的非线性系统可用线性系统近似。

2. 系统分类

按系统与环境的关系,系统可分为孤立系统、封闭系统和开放系统。

1)孤立系统

与外界既没有物质交换,也没有能量交换(如将中性气体放在绝热性很好的、固定的弹性壁做成的匣子内,不计重力)。当外界与系统的相互作用对所研究的问题影响很小时,可看成孤立系统。

2)封闭系统

与外界没有物质交换,只有能量交换的系统称为封闭系统,如在忽略陨石和宇宙尘埃的条件下,地球系统可近似地看作封闭系统。

3)开放系统

与外界既有物质交换,又有能量交换的系统称为开放系统。

另外,按系统与时间的关系,系统可分为静态系统和动态系统。

3. 系统演化的方向

孤立系统演化的惟一方向是平衡态。平衡态是在没有外界影响下,系统内各部分长时间不发生任何变化的状态。平衡态不存任何流和梯度,是最均匀最无序的状态,也是一个没有结构的状态。

封闭系统和开放系统的演化方向是非平衡定态,非平衡定态存在着稳恒的流和梯度。

平衡态是定态的一种特殊情况,平衡态不存在任何流和梯度,如烧热的铁棒退火后所呈现的状态。

4. Benard 对流实验(热驱动)的启示

Benard 对流实验是法国物理学家 Benard 于 1890 年完成的,实验在上下两个恒温的导热板之间充满液体(图 8-1),两板之间的距离 H 远小于板的宽度和长度,上下板的温度分别用 T_1 和 T_2 表示,观察在不同的温度差 ΔT 的情况下,两板之间液体的对流情况。

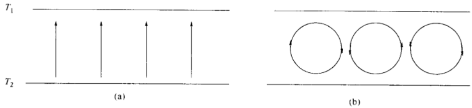


图 8-1 Benard 对流实验

实验观察结果表明:

(1)$\Delta T = 0$ 时,液体处于无序的平衡态;

(2)ΔT 不超过某一阈值 T_{c1} 时,液体在宏观上以静止的热传导方式传递热量,如图 8-1(a)所示;

(3)当 $\Delta T \geqslant T_{c1}$ 时,出现宏观对流态稳定的对流花样,如图 8-1(b)所示,该图形的基本组

成单元是以相反方向旋转的两个流体“卷”；

(4)当 $\Delta T \geqslant T_{c2}$时，出现随时间周期变化的对流花样；

(5)当 $\Delta T \gg T_{c2}$时，出现混沌状态。

二、混沌

1.混沌现象的发现

混沌现象是气象学家 Lorenz 1963 年发现的，他为了预报天气变化，把大气动力学方程简化为 12 个方程用计算机做数值模拟，结果发现在相同的初始条件下，重复模拟结果会随计算时间的增加而彼此分开，最后变得毫无相似之处。Lorenz 用截谱方法把大气动力学方程进一步简化为 3 个一阶的常微分方程组来研究热对流问题，在一定的参数范围内得到了同样的结论，并在三维相空间中观察到混沌行为。

2.混沌定义

混沌是在确定性系统中出现的随机性态，是完全由定律支配的无定律状态。一般认为，确定性系统的长时间行为，可能对初值的微小变化极其敏感，这一现象称为混沌。由于混沌，最初有序的动态系统和非线性系统经过一段时间可转变成完全无组织的状态(图 8－2)，还有另外一种反直觉的现象称为反混沌，某些非常无序的系统自发地“结晶”成为高度有序。

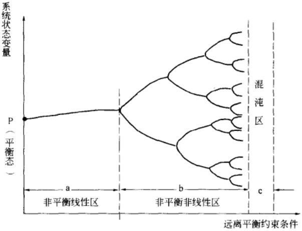


图 8－2　混沌现象

3.混沌是确定论与随机论的统一

人类对自然现象的认识存在两种范式：

(1) 确定论：借助微分方程的高精度分析，它在原理上确定宇宙的整个演化，而实际上只适用于比较简单、结构良好的问题，确定论用于自由度少的简单系统；

(2)平均量的统计分析，它刻画高度复杂系统运动的粗糙特征，统计学则用于自由度多的复杂系统。

在非线性科学看来，秩序不再是定律的同义语，无秩序不再是无规律性的同义语，一种规律属于有秩序性态，而另一种属于无秩序性态，两个范式、两种方法、两种观察世界的方式、两种数学思想体系，各自在它自己的势力范围内适用。

一个确定性系统可产生随机性态，一个随机现象背后可能有确定性定律支配，混沌理论把二者有机地统一起来。

——把湍流考虑成这样的数学对象：一个或几个奇异吸引子也称怪引子，它是流体系统的非常高维数的相空间中较小维数的流形，在此怪引子上的运动，对初值十分敏感，所以流体运动呈现出非周期的似随机特征。这样看来，表面上漫无规则的运动是服从严格的物理定律的，是流体方程的解。

——在热或流体中，虽然系统的每个分子都遵循牛顿力学，这是确定性的，但由系统的微观态到宏观态出现了不确定性，一个平衡态对应许多微观态，这种一对多的对应关系就导致了不确定性，我们不知道系统到底处于哪个确切的微观态，而只知道系统处在每个微观态的几率是多大。对于孤立系统，众多不同的非平衡态最终都要归于一个平衡态，这又是一对多的对应关系，从最终的平衡态无法判断系统来自于哪个非平衡态，这就是普里高津(Prigogine)所认为的，未来并不完全包含在过去之中。

4.混沌不是简单的无序和混乱，而是某种不规则性，更像是没有周期性的次序

混沌更像是没有周期性的次序，在理想模型中，它可能包含着无穷的内在层次，层次之间存在着"自相似性"或"不尽相似"。在观察手段的分辨率不高时，只能看到某一个层次的结构，提高分辨率后，在原来不能识别之处又会出现更小尺度上的结构。完全的决定论和纯粹的概率论，都隐含着承认以某种"无穷"过程为前提。然而，只要承认有限性，那么决定性和概率性描述之间的鸿沟就消失了，决定论的动力学可以产生随机性的演化过程。

总的来说，混沌的产生经历了一个从无序到有序，又从有序到无序的进程(如 Benard 对流实验)，然而混沌不是简单的无序，它和平衡态的无序有本质差别，在混沌无序中还包含着有序，是一种有结构的无序。

三、耗散结构——系统演化的热力学机制

1.热力学定律、熵和熵变

1)热力学定律和熵

(1)热力学第一定律。

热力学第一定律就是能量守恒和转化定律。它是在总结了大量实验工作的基础上得出来的，也是自然界最普遍的规律之一。当系统的状态发生变化时，系统的能量也随之发生变化，这种能量变化的数量关系由热力学第一定律支配。

系统的状态是由状态参量定义的。由状态参量完全确定的函数称为状态函数。状态函数的一个重要性质是：当系统由一个状态变为另一个状态时，状态函数的变化只取决于这种变化的初态和终态。系统的内能就是这样一种状态函数。所谓内能，是指系统内粒子的热运动能量和粒子间相互作用能量的总和。

如对于一个封闭系统，其内能的改变是通过系统对外界作功或系统和外界交换热量来实现的，但只要过程的初态和终态确定，系统内能的变化便完全确定，与作功和热交换的方式无关。相反，即使过程的初态和终态确定，系统对外界作功或系统和外界交换热量都没有确定的值，因为，它们取决于同外界的作用方式，而不是单独由状态参量所能确定的。然而只要过程的初态和终态确定，系统对外界作功和系统与外界交换热量的总和却是确定的值，它正好等于系统内能的变化，即

$$\Delta E = Q - W \qquad (8-4)$$

式中,ΔE 为系统初态到终态的内能增量;Q 为外界传递给系统的热量;W 为系统对外界作的功。

(2)热力学第二定律。

任何不可逆过程都是沿着单向进行的,人们把各种不可逆过程表现出的单向变化的共同规律总结归纳为热力学第二定律,按 Clausius 的说法,不可能把热从低温物体传到高温物体而不产生其它影响。

(3)熵和熵变。

热传导具有不可逆性(单向性),那么过程的初态和终态就存在着相互关系,即不可能用任何办法使终态恢复到初态而不产生其它变化。这样,如果找到一定适当的态函数,就有可能描述不可逆过程的单向性。这个态函数首先由 Clausius 找到,称为熵,用 S 表示。

熵是系统的分子热运动所引起的无序性的一种度量。无序是和大量的各种不同可能性联系在一起的。在无外界影响的情况下,不可逆过程总是沿着从非均匀态向均匀态使熵增加的单一方向进行的。系统宏观状态的均匀程度在微观上反映了分子运动的杂乱程度,即反映了分子运动方式的可能性多少。系统的宏观状态越均匀,它的分子微观热运动越混乱,这种热运动的混乱程度代表着系统状态的无序程度。由此可见,系统的非均匀宏观态的无序程度较低,对应的熵值也较小,而均匀宏观态的无序程度最大,对应的熵值也最大。这种关系首先由 Boltzmann 提出,称为 Boltzmann 原理,其表达式为

$$S = K_b \ln W \tag{8-5}$$

式中,K_b 为 Boltzmann 常数,其值为 $1.38\times10^{-23}\,\mathrm{J\cdot K^{-1}}$;$W$ 为系统宏观状态的热力学几率(系统宏观状态的热力学几率,指系统在任一宏观状态所对应的微观状态数)。

对于只有热接触的封闭系统,一个微过程的熵变 dS 满足 Clausius 不等式

$$dS \geqslant \frac{dQ}{T} \tag{8-6}$$

式中,Q 为外界传递给系统的热量;T 是系统吸收外界热量时的温度。

式(8-6)中的等号对可逆过程成立,不等号对不可逆过程成立。需要注意的是,“可逆系统”只具有数学上的意义,是系统演化的一种极限状态,在系统处于平衡态的时候,我们才可以说系统是可逆系统,自然界的系统皆为不可逆系统。

对于孤立系统,系统与外界无热量交换,即 $dQ=0$,式(8-6)变为

$$dS \geqslant 0 \tag{8-7}$$

它表明孤立系统中的任一不可逆过程的发生都将导致熵的增加。也就是说无论是何种不可逆过程,它都将沿着使熵增加这个单一方向进行,最终达到平衡态,使熵增加到它的最大值。这一表述称为熵的增加原理,也是热力学第二定律的最初数学表述。熵不同于能量,只要不可逆过程发生,它就会被生产出来,而能量不会无中生有,它只能从一种形式转化为另一种形式。

2)热力学第二定律的推广——耗散结构

1956 年,法国物理学家 L.布里渊(L·Brillouin)将信息熵与热力学熵直接联系起来,成功地建立了信息的物理模型和广义增熵原理,并提出了“信息的负熵原理”,即信息与负熵等价。

1969 年,比利时物理化学家普里高津(Prigogine)针对非平衡统计物理学的发展,从热力学第二定律出发,正式创立了一种非平衡态开放系统的自组织理论——耗散结构论。该理论被誉为 70 年代化学的辉煌成就之一,并于 1977 年荣获了诺贝尔化学奖。

普里高津宣称:“非平衡可能成为有序之源,而不可逆过程导致所谓耗散结构这一种新型的物质动态。”

普里高津（Prigogine）等人把热力学第二定律推广到非平衡开放系统，对某一微过程熵变 dS 可以表示为两项之和

$$dS = d_eS + d_iS \tag{8-8}$$

式中，dS 为熵变；d_eS 为熵交焕；d_iS 为熵产生。

d_eS 是由于系统与外界交换物质和能量而引起的熵变，称为熵交换。熵交换可正、可负或为零。

d_iS 是系统内各种不可逆过程所产生出来的熵，熵只能产生不能磨灭，所以熵产生具有非负性，即 $d_iS \geqslant 0$。

既然熵交换 d_eS 可为负值，则式(8－8)中的熵变 dS 也可出现负值，此时，系统朝着熵减小的方向演化，即系统向有序方向演化，或称之为进化，出现的有序结构又称耗散结构。

2.耗散结构(自组织)及其产生的条件

耗散结构是在远离平衡的条件下，借助于外界的能量流和物质流而维持的一种空间或时间的有序结构。

1)退化

孤立系统和处于非平衡线性区的开放系统总是朝着均匀、无序、简单和低级的退化方向演化，好像大自然在企图消灭万事万物的差别，以达到“全球同此凉热”之景观。

2)进化(耗散结构)

在远离平衡的条件下，借助于外界的能量流和物质流而维持的一种空间或时间的有序结构，也可称为合作现象或非平衡非线性现象。

能量流入系统引起正熵流，流出引起负熵流；物质流入系统引起负熵流，流出引起正熵流。

在能量流出和/或物质的流入引起的负熵流大于系统熵产生时，系统向有序方向演化，出现耗散结构，反之，系统向退化方向演化。

3)耗散结构产生的条件

(1)内部条件。

线性系统只存在两种演变前途，不是衰亡就是无限增长，线性系统的定态和非平衡约束是一一对应的，即除热力学分支外无任何其它分支存在。即使外界条件把系统驱动到远离平衡态，甚至把稳定的热力学分支变为不稳定，在不稳定热力学分支附近的扰动也只能无限地发展，因为作为扰动发展归宿的其它稳定分支不存在(图 8－3)。

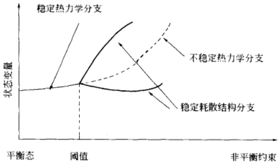


图 8－3　系统演化模式示意图

而非线性系统的演化具有多样性,多样性是系统内部非线性相互作用的结果。在近平衡区,由于非线性项与线性项相比是个小量而不发挥作用,这时只存在热力学分支单解,当非平衡约束把系统驱动到远离平衡态时,非线性项就发挥了主导作用,一个非平衡约束就对应多重定态解,有的定态解是稳定的,有的是不稳定的。在不稳定分支附近的扰动随时间增长,最后演变到某一稳定的分支上,究竟选择哪一个稳定分支是由涨落来确定的(图 8-3)。

非线性相互作用通常表现为正反馈和负反馈,正反馈使系统不稳定但促进系统的发展,负反馈制约正反馈使系统趋于稳定。

(2)外部条件。

在外界驱动下达到非平衡的非线性区才有可能出现耗散结构。如图 8-3 所示,过临界点热力学分支失稳,系统进入稳定的耗散结构分支。

4)涨落导致有序

温度、压强、能量等宏观量都是大量微观粒子行为的统计平均值,而在某一时刻这些宏观量的实际值并不精确地等于这些平均值,或多或少有些偏离,这些偏离就叫涨落,涨落是随机的。

当系统远离平衡态时,一个非平衡约束对应多重定态解,有的定态解是稳定的,有的定态解是不稳定的。在不稳定分支附近的扰动随时间增长,最后演变到某一稳定的分支上,究竟选择哪一种稳定分支是由涨落来完成的。

系统新的稳定有序态是初始偶然的小涨落被选择放大形成的,一旦做出某种选择,其它选择就被排斥,整个过程就不可逆转,原来机会均等的对称性被涨落给破坏,这又被称为对称性破缺。

四、协同学和突变理论

协同学不受热力学约束而从动力角度来研究从无序产生有序结构的规律,其研究内容为:探索在系统宏观状态发生质的改变的转折点附近,支配子系统协同作用的一般性原理。协同学的突出贡献是发现了在分支点附近慢变量支配快变量的普遍原理并以动力学表述,这个原理使对自组织形成的机制有了更深刻的认识。

突变理论是研究系统的状态随外界控制参数的连续改变而发生不连续变化的数学理论,突变现象的共同特点是:外界条件的微变导致系统宏观状态的剧变。突变论为耗散结构论、协同学提供了定量化工具。突变论对拉普拉斯的机械决定论即知道初态就能预知终态的线性思维方式以颠覆性的冲击,把偶然性与必然性、随机性与决定性有机地统一起来,适应研究一切复杂系统的需要。

第二节　沉积盆地内部的复杂结构及其产生的非线性机制

沉积盆地是一个巨型的低温化学反应器,其中发生着一系列的化学反应—物质传输—机械过程,这些过程之间彼此相互影响,具有固有的非线性的本质特征,是一个非线性动力系统。同时,这一非线性动力系统又是开放的,系统与外界之间存在物质交换(沉积物的堆积和剥蚀,流体的流入和流出)和能量交换(构造应力、热流)。

在含油气盆地这样一个极其复杂的系统中可出现宏观上的自组织结构,而正是这一自组织结构控制着油气的生、运、聚、保等过程,决定了油气资源的结构及时空分布规律,认识这一自组织结构的特征,可有效地研究油气藏形成和分布的规律并指导油气勘探与开发。

一、沉积盆地内部的复杂结构

在一个正在沉降和接受充填的沉积盆地中，从浅部到深部一般可分为上、中、下三个次级系统，图 8－4 给出了盆地的这种结构特征。上系统的典型特征是流体压力基本处于正常状态，流体向上自由排出；中系统的典型特征是细粒岩层中的水向上或向下排入到相对高渗透的粗粒岩层中，而流体在后者之中发生连续的侧向流动，在该系统中流体异常压力现象开始出现，且细粒岩层中的压力大于相对高渗透的粗粒岩层中的压力；在下系统中流体流动受到限制，形成流体封存箱，箱内流体呈超压状态，压力梯度在 16MPa/km 和 23MPa/km 之间，且每个流体封存箱内的流体势可以呈均衡状态，处处相等，流体主要沿断裂或裂缝带非常缓慢地连续或呈幕式向上排出，也可突破侧封闭层侧向流动。

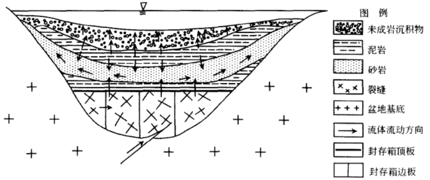


图 8－4　正在沉降和接受充填的沉积盆地内部的一般结构模式

沉积盆地在剖面上的结构性特点实际上反映了盆地随时间的演化特征，而这种结构性特点可用图 8－5 所示的压力—深度曲线来表示。把沉积盆地看作一个天然系统，其中发生着各种物理和化学作用，是一个巨大的化学反应器，盆地深部的流体封存箱是非线性动力系统远离平衡态时所产生的自组织现象。下面，采用非线性动力学的观点，进一步分析从上部到下部盆地系统的演化规律，从而给出关于盆地演化的一个整体认识。

二、沉积盆地内部复杂结构成因的非线性机制

在沉积物埋深不大时，其所处的温度—应力环境与沉积时相差不大，为低能环境，浅层低能环境下，随沉积物的压实，流体可以畅通地排泄，流体压力呈正常状态，系统处于平衡态（图 8－2 中的 p 点），系统中流体柱之间互不相关，处于无序状态，对应于图 8－4 中的上系统。

随着沉积物埋深的增加，沉积物逐渐被压实，并伴随着各种成岩作用而演化成层，渗透性降低，且由于粘土矿物的定向排列作用，使垂向渗透率低于水平渗透率，流体不能及时沿垂向排出，产生弱超压现象，系统偏离平衡态，但处于非平衡线性区（图 8－2 中的 a 区），不产生自组织现象，即不出现侧向流体封闭，形成图 8－4 所示中系统侧向流动的特点。

随着地层埋深的增加，其温度和地应力也在不断增加，沉积物本身、流体本身以及沉积物和流体之间发生的各种搬运、机械变形和化学反应等过程愈趋活跃，这些过程的共同作用使系统远离平衡态（图 8－2 中的 b 区）。在远离平衡态的系统中，出现空间上和时间上的自组织现象，一方面在盆地深部营造出大大小小的高压三维流体封存箱（图 8－4 中的下系统），另一方面，流体从封存箱中的排出在时间上表现为幕式结构。

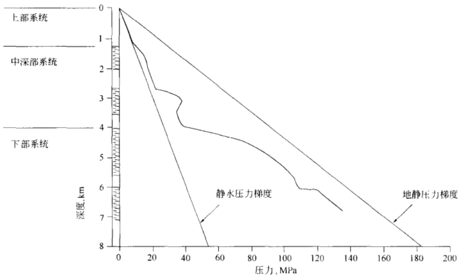


图 8 - 5　对应图 8 - 4 的压力—深度剖面

系统继续演化,更加远离平衡态(图 8 - 2 中的 c 区),则系统在空间和时间上出现的自组织形式就越复杂直至出现混沌,在这种高度演化的系统中,不利于大规模油气藏的形成。

反馈是导致结构性出现的一个重要因素,在盆地中反馈是一个常见现象,故盆地中的空间自组织现象应该是丰富多彩的。

真实的盆地只是可变地、动态地维持在一类远离平衡状态,如沉积停止,盆地边界上的热应力和热作用固定不变,那么盆地演化向岩石零孔隙度、水在上、所有反应处于平衡态。然而,盆地中的沉积、流体输入、剥蚀、应力史非常复杂,在这一历史过程中,盆地可长时间维持在远离平衡态,这期间可出现自组织现象。

含油气盆地是一个多变量、多过程彼此相互制约的复杂系统,不同的变量有不同的变化速率,不同的过程有不同的进程,变量之间以及过程之间可能会出现不协调性,即便系统中的许多过程是快速的和不可逆的,只要有任何一个过程处于平衡态之外,系统则可处于整体上的非平衡态。在远离平衡态的系统中,系统可产生自组织结构(耗散结构),而且,系统出现自组织的可能性随其远离平衡态程度的增加而增加,由此可推断在盆地内出现自组织结构是一个普遍现象。

三、控制沉积盆地内部结构发育的主要因素

图 8 - 4 所示的结构模式是在一个正在沉降和接受充填的沉积盆地内部的一般结构模式,这些结构在什么深度出现,取决于盆地系统在不同深度上的自由能差的大小,即系统远离平衡的程度。在一定深度上系统远离平衡的程度,基本上可由流体超压的大小来衡量,而流体超压的大小主要受盆地的热史、沉积(降)史、构造史的控制。

在一个盆地中,地温梯度越高,沉积速率越大,沉积物中的泥质含量越大,构造挤压越强,则具有侧向流动特点的中系统和高压流体封存箱下系统出现的深度就越小,甚至可由平衡态的上系统直接过渡到下系统的高压流体封存箱,流体封存箱也越复杂(高度进化的结果),表现为大封存箱内套小封存箱的多重巢状结构和/或箱内强烈的非均质性。反之,则中系统和下系

统出现的深度就越大，下系统的流体封存箱也越简单，甚至不出现下系统流体封存箱。

当盆地停止下沉，并出现地层抬升剥蚀，以上的结构模式就可能得到调整。图 8－4 中的上系统变成正常压力状态下垂向重力驱动流系统；中系统在盆地边缘强烈抬升的情况下可能被淡水侵入而变成侧向重力驱动流系统；下系统高压流体封存箱内的压力降低，在一定的条件下可转化为低压流体封存箱。

如何定量地研究盆地的热史、沉积（降）史、构造史对盆地系统结构的控制作用，获得盆地演化阶段的指示性指标（如熵变的方向），是目前地学界遇到的一个挑战，需要多学科联合攻关。

第三节　分形的起源与基本概念

一、自然现象的复杂性和分形的提出

古代的几何学在希腊曾大放异彩，但它研究的图形只是能用圆规和直尺画出来的简单图形，这样的图形全都是平滑的。牛顿微积分学创立后，由微积分学和几何学的结合，使数学能表现更为复杂的形状，但这些几何形状的重要特征是具有特征长度，是平滑的和可微分的。然而，自然现象或自然界出现的某些图形是非常复杂的，如山脉的起伏、海岸线的形态变化等，很难用欧几里得几何或牛顿的微积分学来描述，对于这类复杂现象的描述需要建立新的数学方法和引入新的数学工具。

在复杂的自然现象或几何图形中，有一类具有自相似性或膨胀对称性，即局部与整体在形态、功能和/或信息等方面具有统计意义上的自相似性，自相似性或膨胀对称性又称标度不变性或无标度性，如云彩的外形和海岸线的变化常被当作自然界分形的例子，在不同尺度下观察海岸线，其形态具有相似性，如图 8－6 所示。同时，用不同长度的尺子来丈量海岸线，得到的海岸线长度是不同的，即海岸线具有无标度性。

这类具有自相似性的现象或图形称为分形（fractal）。分形（fractal）一词由法国数学家 Mandelbrot 于 1973 年首次提出，用于描述自然界存在的这类复杂现象。

分形理论属非线性科学的范畴，它已发展成非线性科学中的一个活跃的数学分支，其研究的对象是在非线性系统中产生的不光滑和不可微的自然现象或几何形体，对应的定量参数是分形的维数。分形对微分的否定，既是对微积分学的补充和完善，也是对微积分学的一个划时代的革命。

分形扩大了人类对自然现象研究或描述的范围，原来用欧几里得几何或牛顿的微积分学无法描述的现象，现在可以借助分形几何来研究了，分形为我们提供了描述混沌形态等复杂现象的新语言。

二、分形和非分形的实例

1. 分形的实例

分形是一种复杂的自然属性或几何体，大至宇宙星云的分布，小至准晶态的晶体结构，在地学、生物学、物理学、化学、甚至社会科学中都广泛存在分形现象，有人把这种现象称为分形态或分形体。分形态是介于无序和有序之间的一系列状态，是自然界最广泛分布的状态，这种状态是不稳定的，也可能是准稳定的，这正是自然界运动、发展、演化的根本原因。

地学中分形的例子包括海岸线、山的形态、河流、断裂系统、地震、矿石品位分布等；生物学

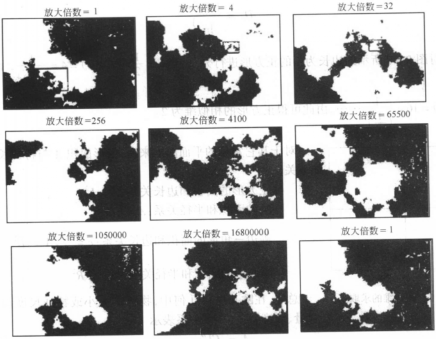


图 8－6　海岸线的自相似性

中的分形包括肺和血管的构造、植物的构造等；宇宙学中的分形包括星云的空间分布、月坑、小行星的直径分布；物理化学中的分形包括固体表面、凝集体、粘性指进、放电式样等。

2. 非分形的实例

尽管，分形在自然界是广泛存在的，但并不是所有的复杂几何形体都是分形，如新的火山锥、冲积扇等自然几何体并不具有分形的特点。

第四节　分形几何——分数维

一、欧几里得几何——拓扑维

1. 拓扑维 D_t

一个几何对象的拓扑维数 D_t 确定了其中一个点所需要的独立坐标数目（经验维）。如在一个长方形中，确定其中一个点的位置，需要两个独立的坐标，故长方形的拓扑维数等于 2。需要注意的是，几何对象的拓扑维与该几何对象所处空间的维数是两个不同的概念。邻域是拓扑学中的一个核心概念，一个几何对象的相邻点，只要保持连续性，无论通过怎样的拉伸、压缩、扭曲变成各种形状，相邻的点仍然相邻，所以说拓扑维是拓扑变换的不变量。

2. 拓扑维的相似法求解——相似维

用标度 r 来分割一个几何对象，得到的分割图形的个数为

$$N(r) = \frac{1}{r^{D_t}} \tag{8-9}$$

则用相似法求得的拓扑维 D_t 为

$$D_t = \frac{\ln N(r)}{\ln \frac{1}{r}} \tag{8-10}$$

如对图 8-7 所示的边长为 1 的正方形进行分割，当 $r = \frac{1}{2}$时，$N(r) = 4 = \frac{1}{(\frac{1}{2})^2}$，而 $r = \frac{1}{4}$ 时，$N(r) = 16 = \frac{1}{(\frac{1}{4})^2}$……，由此可得正方形的相似维为 2。

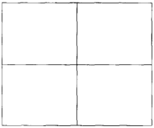


图 8-7　相似维的求取

对于其它常见的平面图形来说，图形面积 A 与特征长度 L 有如下关系

正方形的面积和边长关系：$A = L^2$

圆的面积和半径关系：$A = \pi R^2$

等边三角形的面积和边长关系：$c = \frac{\sqrt{3}}{4}, A = cL^2$

半圆形的面积和半径关系：$A = \frac{\pi}{2} R^2$

总之，在欧几里得几何中，物体的大小或特征尺度与其面积(容量、测度)可用如下关系表示

$$A = CL^{D_t} \tag{8-11}$$

式中，D_t 称为欧几里得空间几何形体的拓扑维数，C 称为标度系数，L 称为特征尺度。

3. 欧几里得几何的特点

(1)维数 D_t 为整数：欧几里得几何研究的图形只是用圆规和规尺画出来的简单图形，这样的图形全都是平滑的，所定义的拓扑维数 D_t 是整数；

(2)$N(r)$随 r 变小而增大，但几何对象的总测度(总长度、总面积或总体积)保持不变。

二、分维数的定义和计算

1. 定义

分维数是分形几何学的特征量，不同的研究人员提出了许多定义，这里仅给出 Kolomogorov 容量维的定义如下

$$D_f = \lim_{r \to 0} \frac{\ln N(r)}{\ln \frac{1}{r}} \tag{8-12}$$

分维 D_f 大于拓扑维 D_t，小于分形几何体所处的空间的维数 D，即 $D_t \leq D_f \leq D$。

2. 分形体实例——科克曲线(Koch 曲线)

图 8-8 展示了 Koch 曲线的产生过程，将这一过程无限地进行下去，获得的复杂几何图形即为 Koch 曲线。



图 8-8　Koch 曲线(科克曲线)

下面，我们根据 Kolomogorov 容量维的定义，来计算 Koch 曲线的分维数。

根据 Koch 曲线的生成过程，我们有

$$\text{当 } r = 1 \text{ 时}, N(r) = 1, L(r) = 1$$

$$\text{当 } r = \frac{1}{3} \text{ 时}, N(r) = 4, L(r) = \frac{4}{3}$$

$$\text{当 } r = (\frac{1}{3})^2 \text{ 时}, N(r) = 16 = 4^2, L(r) = (\frac{4}{3})^2$$

由此可得

$$\text{当 } r = (\frac{1}{3})^n \text{ 时}, N(r) = 4^n, L(r) = (\frac{4}{3})^n$$

根据式(8-3)的定义，可得 Koch 曲线的分维数为

$$D_f = \lim_{r \to 0} \frac{\ln N(r)}{\ln \frac{1}{r}} = \lim_{r \to 0} \frac{n \ln 4}{n \ln 3} = \frac{\ln 4}{\ln 3} = 1.2618$$

另外，也可以发现，当 $n \to \infty$ 时，$L(r) \to \infty$，即当 $r \to 0$ 时，Koch 曲线的总长度 $L(r) \to \infty$。

3. 分形几何与欧几里得几何的对比

表 8-1 归纳出分形几何和欧几里得几何的不同特点。

表 8-1　分形几何与欧几里得几何对比表

欧几里得几何	分形几何
传统的(2000 多年以前)	现代的(近 30 年)
基于特征的大小或尺度	没有具体的大小或尺度
适合于人造的物体	适合于自然界的形态
可用公式描述	用迭代方法产生
整数维	分数维

4. 分维数的计算方法——计盒维数(Box - Counting Method)

用边长为 r 的盒子覆盖图形，数出有图形的盒子数 $N(r)$，变换 r，得出[$N(r)$，$1/r$]的数据对，把[$N(r)$，$1/r$]数据点对投到双对数坐标中，求出直线的斜率，该斜率即为分维数(图8-9)。

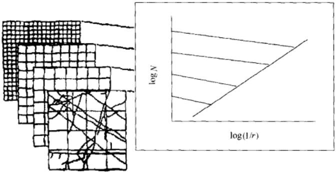


图 8-9　分维数的计盒维数法

第五节　分形问题讨论

一、分形定义的讨论

自 Mandelbrot (1973)给出了分形的定义以后，1986 年又对分形的定义作了修正，修正后的定义为：分形是指由各个部分组成的形态组成的整体，每个部分以某种方式与整体相似。对这一定义加以引申，它包括如下含义：

(1)分形既可以是几何图形，也可以是由“功能”或“信息”构建的数理模型；

(2)分形可以同时具有形态、功能和信息等多方面的自相似性，也可以只有其中某一方面的自相似性；

(3)自相似可以是严格的，也可以是统计意义上的相似，自然界的大部分分形都是统计自相似，而不是局域放大一定倍数以后，简单地与整体完全重合，但是分形维数并不会因放大或缩小等操作而变化(即分形体具有伸缩对称性)；

(4)相似性有层次结构上的差异，数学中的分形具有无限嵌套的层次结构，而自然界中的分形只有有限层次的嵌套，且要进入到一定的层次结构以后才有分形的规律；

(5)相似性有级别(即使用生成元的次数或放大倍数)上的差异，级别最高的是整体，最低的称为 0 级生成元，级别愈接近则愈相似，级别相差越大，相似性愈差。相似性级别可用无标度区间或标度不变性范围表示。

二、分形与复杂系统

传统科学把观察到的无序状态作为有序系统的一种摄动，认为无序系统继承了有序系统的物理或化学实质，与有序系统的偏差是很小的，但现代非线性系统理论告诉我们，如果无序性很强，我们不能指望从有序系统理论了解系统的更多信息。分形理论恰恰适应于无序系统，它把无序性看成是系统固有的性态，而不是把它看成摄动。在区分强无序和弱无序时，其效果也十分明显。

说一个系统是弱无序的，如果我们用逐渐减小的尺度(对连续模型)或用逐渐增大的尺度(对点阵模型)来考察或探测某一系统时，无序性消失。如水晶体表面和玻璃，在小尺度或大尺度下，无序性与平移不变参考态有确定的偏离。强无序指在小尺度或大尺度下，无序性仍然保持且不存在可与系统比较的参考态。因此，无序性在所有标度上的重复，对应于分形几何中结构之内结构的重复性，结果它的几何形状就不同于经典几何图，即它是分形。分形维数与拓扑维数的差值 $D_f - D_t$ 是系统无序程度的量度之一。$D_f = D_t$ 的大多数系统是有序或弱无序的，$D_f > D_t$ 的系统是强无序的。

分维数是粗糙度(不平度)的一种度量，是物体占据空间能力的一种量度。分形是自然界中普遍存在的现象，但并不是所有的自然现象都是分形结构，有两方面的解释：一是从分形形成的机制上来看，有些过程不产生分形；二是从无标度区间的范围来看，在无标度区间之外，自相似性消失或不存在。

三、关于分形分布和自相似性

分形分布要求大于某一尺度物体的数目与物体大小之间存在幂函数关系

$$P\{Z > z\} \propto z^{-D_f} \qquad (8-13)$$

关于地质现象的统计分布,幂函数分布不是惟一的分布,但幂指数分布是惟一不含特征长度的分布。分形系统在伸缩变换下保持不变且空间坐标 x、y 和 z 按相同比例伸缩(各向同性无序),更一般的自仿射系统在不同的方向上有不同的伸缩比例。有序系统也是伸缩不变的,而且它还具有平移不变性,而分形则失去了平移对称而保持了伸缩对称。事实上,每种无序都以某种方式破坏了平移对称性,这说明具有整数维 1、2、3 的系统是复杂分形系统退化的,而非整数维系统没有这种退化。

以上关于自相似性的定义是从数学角度出发的,从物理学角度出发,自相似性是指某种结构或过程的特征从不同的空间尺度或时间尺度来看都是相似的,或者某系统或结构的局域性质或局域结构与整体类似。自相似性是标度不变性的同义语,它强调系统在各种尺度上是相似的这一事实。

第六节　分形在油气勘探开发中的应用

近几年来,国内外许多专家,如 Bale(1984)、Katz(1985)、Krohn(1986)、Avnir(1985)及沈平平(1994)和李克文等(1993)发表了他们对储层非均质性、砂岩等储集体中孔隙结构等问题进行的大量研究结果,这些研究发现储层的非均质性及多孔介质的孔隙结构是一种分形结构。尽管他们测定或计算分形维数的方法不同,他们都用分形维数描述各自的研究对象,取得了较好的应用效果。下边就分形理论在油气储层勘探开发和评价中的应用介绍其基本原理和实现过程。

一、分形方法在岩石孔隙结构研究中的应用

1.用图像分析技术研究孔隙结构的分形维数

该方法是根据测度关系求取薄片岩样的分形维数(李克文,1993),具体作法如下:

(1)将洗油烘干的薄片样品在高压下灌注染色环氧树脂,冷却后磨成薄片;

(2)在一定倍数的显微镜下将薄片拍成彩色照片,再将彩色照片输入计算机,获得彩色图像,提取反映孔隙特征的二值图像,用图像分析软件进行图像分析;

(3)在处理好的二值图像上,随机地选取一个点作圆心,用一半径画圆,数出该圆内孔隙所占的点数 $N(r)$;

(4)作出不同的 r 与相应的 $N(r)$之间的关系,若 $N(r)$与 r 间满足如下幂指数关系,即 $N(r) \propto r^{D_f}$,则在对数坐标上直线的斜率即是孔隙分布的分形维数。

2.利用压汞毛管压力曲线研究孔隙结构的分形维数

利用毛管压力曲线计算孔隙结构的分形维数是一种新型的分维计算方法(沈平平,1994),该方法是根据压汞获得的毛管压力 P_c 与砂岩样品中进汞饱和度的 S_{Hg}之间的关系,求得进汞饱和度与孔隙喉道半径 r 之间的关系,然后确定在任意给定的孔隙喉道半径 r_c 时,计算所有大于 r_c 的喉道数,统计大于 r_c 喉道出现的概率 $P(r > r_c)$,利用分形理论求得分形维数。具体操作步骤如下:

(1)对砂岩样品进行压汞实验,绘制毛管压力 P_c 与进汞饱和度 S_{Hg}之间的关系;

(2)将毛管压力 P_c 按照著名的 Washburn 方程换算成喉道半径

$$r = \frac{2\sigma\cos\theta}{P_c} \tag{8-14}$$

式中，σ 是水银表面张力，θ 是水银与岩石的接触角。

(3)根据 S_{Hg}—r 的关系，计算所有大于某一随机给定 r_c 的喉道数，统计大于 r_c 喉道出现的概率 $P(r > r_c)$；

(4)在双对数坐标纸上，绘制喉道出现的概率 $P(r > r_c)$ 与 r 的关系曲线，若为直线，则直线的斜率即为分形维数。

二、分形方法在储层的非均质性研究中的应用

1.分形理论井间变量插值的应用

自从 Hewett(1986)发表了用分形理论描述和评价储层非均质的文章后，分形理论很快在石油工业界受到了重视，其中很重要的原因之一是目前缺乏可用的储层井间参数的预测和描述方法。

分形插值提供了在只有两口井资料的情况下，对井间的变量值进行插值的方法，在只有两口井的情况下(图 8-10)，第六章中的克立金方程组可写成

$$\begin{cases} \lambda_1\gamma(0,l) + \mu = \gamma(l,x) \\ \lambda_2\gamma(0,l) + \mu = \gamma(0,x) \\ \lambda_1 + \lambda_2 = 1 \end{cases} \tag{8-15}$$

在非均质多元介质中，物体运移取决于渗透性分布的空间相关性，沉积环境中物性参数测量结果分析表明在长距离相关中存在分形特点。具有分形特点的变量，其变差函数可由下式表示：

$$\gamma(h) = \frac{1}{2}E[Z(x) - Z(x+h)]^2 = \frac{1}{2}V_H h^{2H} \tag{8-16}$$

式中，V_H 为系数项；H 为 Hurst 指数。

Hurst 指数由以下关系式求取

$$H = D_t - D_f \tag{8-17}$$

式中，D_t 为拓扑维数；D_f 为分维数。

将 $\gamma(h) = \frac{1}{2}V_H h^{2H}$ 代入线性插值公式求解得

$$z^*(x) = \frac{1}{2}[(1-w)z(0) + (1+w)z(h)] \tag{8-18}$$

式中，$w = \left(\frac{x}{l}\right)^{2H} - \left(1 - \frac{x}{l}\right)^{2H}$。

2.分形维数与储层分类

对国内多个油田的砂岩样品的孔隙结构的分析统计表明，砂岩样品具有明显的分形结构，而且所测的分形维数在 2.40～2.97 之间，图 8-11 是国内 6 个油田 178 块岩样测得的分形维数分布(沈平平，1982)。从统计结果看，分形维数愈小，则孔隙均匀，储集性好；反之，则孔隙不均匀，储集性不好。

图 8-12 给出了从某岩样铸体薄片得到的孔隙结构图像，图中的黑色部分是充满铸体的孔隙空间。从图中可以看到，该岩样孔隙结构连通性好，粘土含量小，是第一类孔隙结构的储

层，其对应的孔隙分维数在2.4左右(图8-13)。

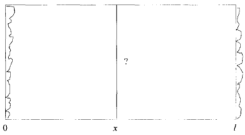


图8-10 已知两口井的变量值(如孔隙度测井数据)，进行井间插值的示意图

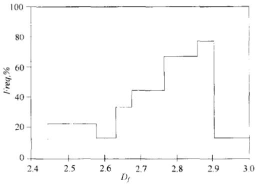


图8-11 国内6个油田178块岩样测得的分形维数分布(据沈平平，1986)

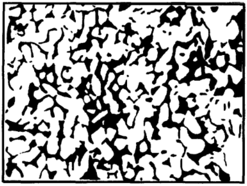


图8-12 铸体薄片的孔隙结构图像

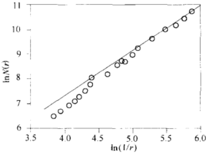


图8-13 图8-12对应的孔隙分形特征

图8-14给出了从另一个岩样铸体薄片得到的孔隙结构图像，该岩样孔隙结构好，粘土含量有所增加，孔隙的连通性差，是第二类孔隙结构的储层，测出的分形维数为2.5~2.6。

图8-15给出了从又一个岩样铸体薄片得到的孔隙结构图像，该岩样大颗粒之间的孔隙被小颗粒及胶结物充填，粘土含量大大增加，孔隙结构相当复杂。总体上看，孔隙少而小，孔隙分布极不均匀，是第四类储层，测出的分形维数为2.9以上。

综上所述，用砂岩储层的孔隙结构的分形维数可以进行储层分类，其描述结果与微观均质系数的分类和油田水驱开发的实际效果吻合较好，可以作为评价油藏孔隙结构和储层性能的一个主要指标。

3. 多段分形特征与储层孔隙结构

从现有的文献看，采用分形方法描述岩石孔隙结构特征时，一般只考虑孔隙结构的线性或一小段的分形分布，如图8-16所示的线性段。然而，实际工作发现，由于储层孔隙结构的复杂性，实测的毛管压力曲线往往表现出多个分段的特征，如图8-17所示。

从图8-16可以看到，分形曲线可以分成喉道半径小于某个值 r_c 和大于该值的两个部分。根据分形理论，对应于直线部分的孔隙具有自相似性特征，称为分形孔隙，而对应于非直

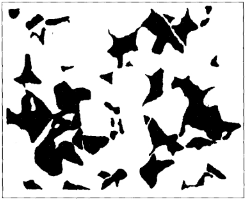


图 8-14 二类储层对应的孔隙结构特征



图 8-15 四类储层对应的孔隙结构特征

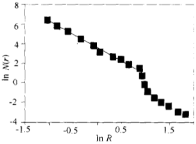


图 8-16 分形分布的线性段特征

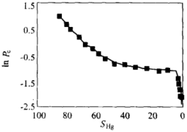


图 8-17 实测毛管压力曲线的多段分布特征

线部分的孔隙称为非分形孔隙。由于沉积环境和成岩作用的不同,孔隙结构表现出多种多样的特征也是自然现象复杂性的表现之一。

图 8-18 给出了从某油田砂岩样品测得的多段分形特征,反映了该区存在三类孔隙分布,对应于储层的不同地质特征。孔隙平均半径大于 r_0 的区域为Ⅲ区,尽管也表现为直线分布但

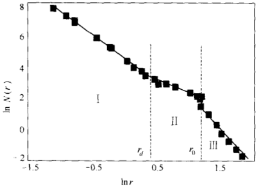


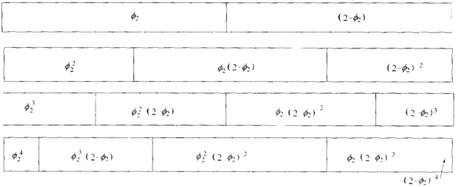
图 8-18 某油田砂岩样品测得的多段分形特征

与其它两个区域的孔隙分布特征明显不同;孔隙平均半径大于 r_d 小于 r_0 的区域为Ⅱ区，孔隙平均半径小于 r_d 的区域为Ⅰ区,从薄片分析资料看,孔隙平均半径小于 r_0 的两个区域都表现为分形结构,但具有不同的分形维数,Ⅱ区称为分形大孔隙区域,而Ⅰ区称为分形小孔隙区域,这两类孔隙区域与开发时的无水采收期密切相关。分维大的区域标志着非均质性强,水驱推进不均匀,因此,无水采收期短。

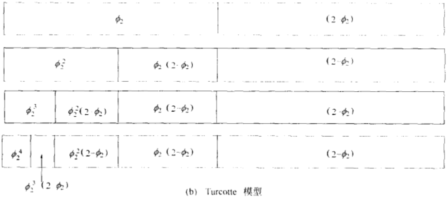
三、分形在油气藏规模研究中的应用

1.成藏(富集)过程

在研究矿产富集规律时,南非矿业统计学家 De Wijis 提出一个服从对数正态分布的富集模型(图 8-19a)。在该模型中,岩体先分成相等的两半,再分成相等的 4 份,如此进行下去以至无穷,而矿物越来越集中在越来越小的体积中。图 8-19a 给出了三级成矿过程,在每一级成矿过程中,原胞的一半按富集因子 ϕ_2 变得富集,而另一半按因子$(2-\phi_2)$变得贫乏。这个模型给出的储量与品位间的关系是二项式分布,当分割体积非常小时,将会得到两者之间的对数正态分布。在 De Wijis 模型思路的启发下,Turcotte 提出一个服从幂函数分布的富集模型,见图 8-19(b)。在 Turcotte 模型中,岩体的分法与 De Wijis 模型相同,但进一步的成矿浓缩限制于最高品位的矿石之中,这将导致储量与品位间出现幂函数(分形)分布关系。



(a) De Wijis 模型



(b) Turcotte 模型

图 8-19 矿产富集的 De Wijis 模型和 Turcotte 模型

De Wijis 模型是在研究了矿产资源量与矿石品位之间统计关系的基础上建立起来的，难以反映成矿机理。和 De Wijis 模型相比，Turcotte 模型强调进一步的成矿浓缩限制于最高品位的矿石之中，从直观上来看这似乎比较确切地反映了成矿的规律，而且 Turcotte 模型是建立在一定的理论基础之上的。在固体矿产资源预测中，分形模型越来越广泛地被应用，也从另一方面说明了 Turcotte 模型的合理性。

2.油气藏规模分布

在图 8-4 的中系统中形成的油气藏群，在成因上相互联系，具有上述 Turcotte 过程或多重 Turcotte 过程的特点，油气藏规模分布服从分形分布是有成因依据的。

在图 8-4 的上系统中形成的油气藏群，在成藏的静态因素上有联系，而在成藏的动态因素上没有联系，油气藏可能是深部油气向上运移充注局部有利圈闭形成的，也可能是由原生的低熟油气在局部有利条件下形成的，不具备 Turcotte 过程的特点，每个油气藏的规模受局部条件控制而相互之间相对独立，从概率意义上来看上系统中的油气藏群的规模分布倾向于服从正态分布。

在图 8-4 的下系统中形成的油气藏群，油气藏的规模主要受圈闭容量（流体封存箱的大小）的控制，而深部流体封存箱是由于系统远离平衡态所形成的自组织结构，这种自组织结构可能具有分形特征，也可能不具有分形特征（Ortoleva，1994），其规模分布可能出现分形分布、对数正态分布及正态分布等不同的情形，导致油气藏的规模分布可能出现多种情形。

从概率论和统计熵的观点来看，从分形分布到对数正态分布，从对数正态分布到正态分布，从正态分布到均匀分布，代表着系统增熵退化（系统抬升，由封闭变为垂向上开启）或者系统高度进化（系统埋深增加）至出现混沌态的两种演化趋势。从分形分布到对数正态分布，从对数正态分布到正态分布，从正态分布到均匀分布，系统的含油气性由好到差，均匀分布是一种极限情况，代表系统中没有形成规模性油气藏（图 8-20）。

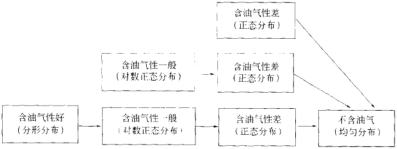


图 8-20 油气成藏流体动力系统含油气性的演化模式

现今存在的油气藏群，其规模的分布既受油气藏群形成时的规模分布控制，又受后期演化程度的控制，不同的规模分布率都可能出现，但富含油气盆地内中系统中的油气藏群，更可能具有分形分布结构，而上系统中的油气藏群更可能具有正态分布结构，下系统中的油气藏群规模分布情况就要复杂得多。总之，在不同的石油地质条件下，可能表现为不同的规模分布率。

大量研究也发现，在勘探程度比较高的盆地中，单井累积探明平均油气储量与勘探井数之间的关系可由下式表示

$$N_p \sim W^{-D_f} \qquad (8-19)$$

式中　N_p——单井累积探明平均油气储量,Mm^3/井;

W——所钻探井的数量;

D_f——分维数。

根据历史数据,可求出分维数 D_f,并可根据以上关系式对未来钻探所发现的储量进行预测。如 Drew&Schuenemeyer 对 1945~1978 年间 Louisana 和 Texas 的勘探资料进行了分析,得到 $D_f=2.22$(图 8-21)。

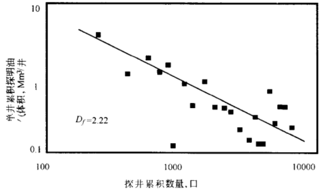


图 8-21　1945~1978 年间 Louisana 和 Texas 单井探明平均油气储量与钻井数的关系

在已知总的储量基础上,对某一勘探区的最终储量进行估计

$$R = N_p\left(\frac{W}{W_0}\right)^{D_f} \tag{8-20}$$

式中　R——单井最终可发现储量,Mm^3/井;

N_p——每口井量积发现储量,Mm^3/井;

W——所钻探井的数量;

W_0——完全勘探一个含油气区所需的探井总数。

公式(8-20)中的 W_0 可由下式估计

$$W_0 = \frac{A_0}{A} \tag{8-21}$$

式中　A_0——勘探区总面积,km^2;

A——给定规模区的已发现油气藏的平均面积,km^2。

四、分形应用实例分析

【例 8-1】　在中国东部某凹陷内有 2 个构造层(T_6 和 T_2),每层又可根据其断层分布特点划分为南北 2 个区,后再划分为东西 2 个区,2 个构造层(T_6 和 T_2)共划分为 8 个区,对每个区采用 0.5km、1km、1.5km、2km 和 4km 的 5 种线性尺度(L_i)分别进行 8 个区中的断层网格覆盖,得到相应的 5 个含断层的网格数目(N_i)如表 8-2 所示,试利用表中数据求分形维 D、线性尺度大小与含断层网格数的相关系数 γ_{LN} 并进行计算结果的地质分析。

解:第一步,利用计盒维数法统计每个线性尺度的网格所含断层的网格数,结果如表 8-2 中的第 3~7 列所示;

表 8-2 T_6 和 T_2 构造层断层系的分形数据及计算结果

构造层	分形对象	线性尺度大小					计算结果	
		$L_1$0.5 km	$L_2$1.0 km	$L_3$1.5 km	$L_4$2.0 km	$L_4$5 km	D	γ_{LN}
T_6	S	1689	782	534	380	167	1.1034	0.99923
	N	1179	560	353	255	95	1.2023	0.99925
	E	1569	719	468	352	144	1.1363	0.99893
	W	1305	613	419	283	118	1.1496	0.99804
T_2	S	1409	715	426	300	149	1.0880	0.99885
	N	930	442	273	196	92	1.1189	0.99986
	E	1453	703	441	319	141	1.1232	0.99961
	W	886	454	256	177	100	1.0812	0.99549

第二步,利用计盒维数中维数的如下计算公式计算每个区域的分形维数，结果如表 8-2 中的第 8 列所示；

$$Dim_B(F) = -\frac{\ln[N_\delta(F)]}{\ln\delta}$$

第三步,利用相关系数计算公式计算每个区域的含断层的网格数与线性尺度的相关系数,结果如表 8-2 中的第 9 列所示；

第四步,对结果的地质分析:(1) 从表 8-2 中列出的相关系数可以看到 $\ln L$ 和 $\ln N$ 的线性相关度较高,即所研究的断层分布符合幂指数分布规律,用分形维数 D 值可定量描述该凹陷的断层特征;(2) 对 T_6 构造层,计算得到的分形维数有 $D_N > D_S$ 和 $D_W > D_E$,而对于 T_2 层构造层,分形维数有 $D_N > D_S$, $D_E > D_W$。从分形维数与几何分布的复杂性关系知道,分形维数愈大,几何图形愈复杂。对应于这一例子,分形维数的大小标志着断层的发育程度。

表 8-2 的结果表明该凹陷构造运动导致断层发育程度为:北高南低。但两个不同构造层东西两区断层发育程度不同。由于断层在该区对油气运移聚集有重要作用,所以根据计算得到的分维值,预测该凹陷油气发育程度为北高南低。根据实际的油气生产情况证实,该凹陷的油藏储量规模和产量都支持了这个计算分析结果。

◆ 思 考 和 练 习 题 ◆

1. 如何理解线性系统和非解线性系统?
2. 按照系统与外界的关系,系统有哪些分类?
3. 如何理解混沌?
4. 什么是非欧氏几何? 与欧氏几何有何区别?
5. 试分析如何用压汞分析资料计算孔隙结构的分形维数?
6. 试用计盒维数法计算斯宾斯基地毯的分形维数。
7. 举例说明自然界中有哪些是分形现象? 为什么说是分形体?
8. 除了研究孔隙结构外,分形理论还可在油气勘探开发的哪些方面应用?

附　表

附表 1　标准正态分布

本表给出了 $F(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt (z \geqslant 0)$ 的数值。

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0.1	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0.2	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0.3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0.4	65542	65910	66276	66640	67033	67364	67724	68082	68439	68793
0.5	69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0.6	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0.7	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0.8	78814	79103	79389	79679	79955	80234	80511	80785	81057	81327
0.9	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1.0	84134	84375	84614	84850	85083	85314	85543	85769	85993	86214
1.1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1.2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1.3	90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1.4	91924	92073	92220	92364	92507	92647	92786	92922	93056	93189
1.5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1.6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1.7	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1.8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1.9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2.0	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2.1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98547
2.2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2.3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2.4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2.5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2.6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2.7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2.8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2.9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
z	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
3	0.99865	99903	99931	99952	99966	99977	99984	99989	99993	99995

附表2 t 分布

本表给出了满足 $P(|t| > t_a) = a$ 的 t_a 值(n 是自由度)。

n	a				
	0.10	0.05	0.02	0.01	0.001
	t_a				
1	6.31	12.71	31.82	63.66	636.62
2	2.92	4.30	6.97	9.93	31.60
3	2.35	3.18	4.54	5.84	12.94
4	2.13	2.78	3.75	4.60	8.61
5	2.02	2.57	3.37	4.03	6.86
6	1.94	2.45	3.14	3.71	1.96
7	1.90	2.37	3.00	3.50	5.41
8	1.86	2.31	2.90	3.36	5.04
9	1.83	2.26	2.82	3.25	4.78
10	1.81	2.23	2.76	3.17	4.59
11	1.80	2.20	2.72	3.11	4.44
12	1.78	2.18	2.68	3.06	4.32
13	1.77	2.16	2.65	3.01	4.22
14	1.76	2.15	2.62	2.98	4.14
15	1.75	2.13	2.60	2.95	4.07
16	1.75	2.12	2.58	2.92	4.02
17	1.74	2.11	2.57	2.90	3.97
18	1.73	2.10	2.55	2.88	3.92
19	1.73	2.09	2.54	2.86	3.88
20	1.73	2.09	2.53	2.85	3.85
21	1.72	2.08	2.52	2.83	3.82
22	1.72	2.07	2.51	2.82	3.79
23	1.71	2.07	2.50	2.81	3.77
24	1.71	2.06	2.49	2.80	3.75
25	1.71	2.06	2.48	2.79	3.73
26	1.71	2.06	2.48	2.78	3.71
27	1.70	2.05	2.47	2.77	3.69
28	1.70	2.05	2.47	2.76	3.67
29	1.70	2.04	2.46	2.76	3.66
30	1.70	2.04	2.46	2.75	3.65
40	1.68	2.02	2.42	2.70	3.55
60	1.67	2.00	2.39	2.66	3.46
120	1.66	1.98	2.36	2.62	3.37
∞	1.65	1.96	2.33	2.58	3.29

附表3 χ^2 分布

本表给出了满足 $P\{\chi^2(n) > \chi^2_\alpha(n)\} = \alpha$ 的 $\chi^2(n,\alpha)$ 值。

n	$\alpha=0.995$	0.99	0.975	0.95	0.90	0.75
1	—	—	0.001	0.004	0.016	0.102
2	0.010	0.020	0.051	0.103	0.211	0.575
3	0.072	0.115	0.216	0.352	0.584	1.213
4	0.207	0.297	0.484	0.711	1.064	1.923
5	0.412	0.554	0.831	1.145	1.610	2.675
6	0.676	0.872	1.237	1.635	2.204	3.455
7	0.989	1.239	1.690	2.167	2.833	4.255
8	1.344	1.646	2.180	2.733	3.490	5.071
9	1.735	2.088	2.700	3.325	4.168	5.899
10	2.156	2.558	3.247	3.940	4.865	6.737
11	2.603	3.053	3.816	4.575	5.578	7.584
12	3.074	3.571	4.404	5.226	6.304	8.438
13	3.565	4.107	5.009	5.892	7.042	9.299
14	4.075	4.660	5.629	6.571	7.790	10.165
15	4.601	5.229	6.262	7.261	8.547	11.037
16	5.142	5.812	6.908	7.962	9.312	11.912
17	5.697	6.408	7.564	8.672	10.085	12.792
18	6.265	7.015	8.231	9.390	10.865	13.675
19	6.844	7.633	8.907	10.117	11.651	14.562
20	7.434	8.260	9.591	10.851	12.553	15.452
21	8.034	8.897	10.283	11.591	13.240	16.344
22	8.643	9.542	10.982	12.338	14.042	17.240
23	9.260	10.196	11.689	13.091	14.848	18.137
24	9.886	10.856	12.401	13.848	15.659	19.073
25	10.520	11.524	13.120	14.611	16.473	19.939
26	11.160	12.198	13.844	15.379	17.292	20.843
27	11.808	12.879	14.573	16.151	18.114	21.749
28	12.461	13.565	15.308	16.928	18.939	22.657
29	13.121	14.257	16.047	17.708	19.768	23.567
30	13.787	14.954	16.791	18.493	20.599	24.478
31	14.458	15.655	17.539	19.281	21.434	25.390
32	15.134	16.362	18.291	20.072	22.271	26.304
33	15.815	17.074	19.047	20.867	23.110	27.219
34	16.501	17.789	19.806	21.664	23.952	28.136
35	17.192	18.509	20.569	22.465	24.797	29.054
36	17.887	19.233	21.336	23.269	25.643	29.973
37	18.586	19.960	22.106	24.075	26.492	30.893
38	19.289	20.691	22.878	24.884	27.343	31.815
39	19.996	21.426	23.654	25.695	28.196	32.737
40	20.707	22.164	24.433	26.509	29.051	33.660
41	21.421	22.906	25.215	27.326	29.907	34.585
42	22.138	23.650	25.999	28.144	30.765	35.510
43	22.859	24.398	26.785	28.965	31.625	36.436
44	23.584	25.148	27.575	29.787	32.487	37.363
45	24.311	25.901	28.366	30.612	33.350	38.291

续表

n	$\alpha=0.25$	0.10	0.05	0.025	0.01	0.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750
6	7.841	10.645	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.194	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993
29	33.711	39.087	42.557	45.722	49.588	52.336
30	34.800	40.256	43.773	46.979	50.892	53.672
31	35.887	41.422	44.985	48.232	52.191	55.003
32	36.973	42.585	46.194	49.480	53.486	56.328
33	38.058	43.745	47.400	50.725	54.776	57.648
34	39.141	44.903	48.602	51.966	56.061	58.964
35	40.223	46.059	49.802	53.203	57.342	60.275
36	41.304	47.212	50.998	54.437	58.619	61.581
37	42.383	48.363	52.192	55.668	59.892	62.883
38	43.462	49.513	53.384	56.896	61.162	64.181
39	44.539	50.660	54.572	58.120	62.428	65.476
40	45.616	51.805	55.758	59.342	63.691	66.766
41	46.692	52.949	56.942	60.561	64.950	68.053
42	47.766	54.090	58.124	61.777	66.206	69.336
43	48.840	55.230	59.304	62.990	67.459	70.616
44	49.913	56.369	60.481	64.201	68.710	71.893
45	50.985	57.505	61.656	65.410	69.957	73.166

附表4 F 分布

本表给出了 $P\{F(n_1,n_2) > F_\alpha(n_1,n_2)\} = \alpha$ 时 $F_\alpha(n_1,n_2)$ 的值。

$\alpha = 0.1$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	68.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	8.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.60	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	8.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2,25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.96	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.93	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.4	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	1.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1 781	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	11.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

续表

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
										$\alpha = 0.05$									
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.58	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	12.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.011	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.08	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.08	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
80	4.00	3.16	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.88	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

续表

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	$\alpha = 0.025$																		
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.62	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
80	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

续表

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	$\alpha = 0.01$																		
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	80.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	18.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5,28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	15.35	5.20	5.11	4.96	4.81	4.73	4.85	4.57	4.48	4.40	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	4.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	3.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
80	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

续表

n_2 \ n_1 ($\alpha = 0.005$)	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.63	12.40	12.27	12.14
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.65	7.58	7.42	7.31	7.19	7.08
8	14.69	11.04	9.60	8.81	8.80	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
11	13.23	8.91	7.60	6.88	6.87	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23
12	11.75	8.51	7.28	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.15	3.06	2.95	2.84	2.73	2.61
22	9.73	6.81	5.65	5.02	4.61	4.39	4.11	3.94	3.81	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
80	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

续表

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	$\alpha = 0.001$																		
1	4053*	5000*	5405*	5625*	5764*	5859*	5929*	5981*	6023*	6056*	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
2	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4	999.4	999.4	999.4	999.5	999.5	999.5	999.5	999.5	999.5
3	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2	128.3	127.4	126.4	125.9	125.4	125.0	124.5	124.0	123.5
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.76	46.10	45.77	45.43	45.09	44.75	44.40	44.05
5	47.18	37.12	33.20	31.09	29.75	28.84	28.16	27.64	27.24	26.92	26.42	25.91	25.39	25.14	24.87	24.60	24.33	24.06	23.79
6	35.51	27.00	23.70	21.92	20.81	20.03	19.46	19.03	18.69	18.41	17.99	17.56	17.12	16.89	16.67	16.44	16.21	15.99	15.75
7	29.25	21.69	18.77	17.19	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.32	12.93	12.73	12.53	12.33	12.12	11.91	11.70
8	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.04	11.77	11.54	11.19	10.84	10.48	10.30	10.11	9.92	9.73	9.53	9.33
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.24	8.90	8.72	8.55	8.37	8.19	8.00	7.81
10	21.04	14.91	12.55	11.28	10.48	9.92	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.64	7.47	7.30	7.12	6.94	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.32	7.01	6.85	6.68	6.52	6.35	6.17	6.00
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.25	6.09	5.93	5.76	5.59	5.42
13	17.81	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.23	5.93	5.78	5.63	5.47	5.30	5.14	4.97
14	17.14	11.78	9.73	8.62	7.92	7.43	7.08	6.80	6.58	6.40	6.13	5.85	5.56	5.41	5.25	5.10	4.94	4.77	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	5.10	4.95	4.80	4.64	4.47	4.31
16	16.12	10.97	9.00	7.94	7.27	6.81	6.46	6.19	5.98	5.81	5.55	5.27	4.99	4.85	4.70	4.54	4.39	4.23	4.06
17	15.72	10.66	8,73	7.68	7.02	7.56	6.22	5.96	5.75	5.58	5.32	5.05	4.78	4.63	4.48	4.33	4.18	4.02	3.85
18	15.38	10.39	8.49	7.46	6.81	6135	6.02	5.76	5.56	5.39	5.13	4.87	4.59	4.45	4.30	4.15	4.00	3.84	3.67
19	15.08	10.16	8.28	7.26	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.70	4.43	4.29	4.14	3.99	3.84	3.68	3.51
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.56	4.29	4.15	4.00	3.86	3.70	3.54	3.38
21	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.44	4.17	4.03	3.88	3.74	3.58	3.42	3.26
22	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.33	4.06	3.92	3.78	3.63	3.48	3.32	3.15
23	14.19	9.47	7.67	6.69	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.23	3.96	3.82	3.68	3.53	3.38	3.22	3.05
24	14,03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.14	3.87	3.74	3.59	3.45	3.29	3.14	2.97
25	13.88	9.22	7.45	6.49	5.88	5.46	5.15	4.91	4.71	4.56	4.31	4.06	3.79	3.66	3.52	3.37	3.22	3.06	2.89
26	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	3.99	3.72	3.59	3.44	3.30	3.15	2.99	2.82
27	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.92	3.66	3.52	3.38	3.23	3.08	2.92	2.75
28	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.36	4.11	3.86	3.60	3.46	3.32	3.18	3.02	2.86	2.69
29	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.80	3.54	3.41	3.27	3.12	2.97	2.81	2.64
30	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.75	3.49	3.36	3.22	3.07	2.92	2.76	2.59
40	12.61	8.25	6.60	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.40	3,15	3.01	2.87	2.73	2.57	2.41	2.23
80	11.97	7.76	6.17	5.31	4.76	4.37	4.09	3.87	3.69	3.54	3.31	3.08	2.83	2.69	2.55	2.41	2.25	2.08	1.89
120	11.38	7.32	5.79	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.78	2.53	2.40	2.26	2.11	1.95	1.76	1.54
∞	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	2.13	1.99	1.84	1.66	1.45	1.00

注：* 表示要将此数乘以 100

参 考 文 献

[1] 傅民禄. 数理统计在两江铜矿中的应用. 桂林冶金地质学院学报, 第 1 期, 1981
[2] 方开泰, 潘恩沛. 聚类分析. 地质出版社, 1982
[3] 刘承祚, 孙惠文. 数学地质基本方法及应用. 地质出版社, 1982
[4] 王学仁. 地质数据的多变量统计分析. 科学出版社, 1982
[5] 方开泰, 潘恩沛. 聚类分析. 地质出版社, 1982
[6] 王学仁. 地质数据的多变矿床统计预测. 地质出版社, 1982
[7] 赵鹏大, 胡旺亮, 李紫金. 矿床统计预测. 地质出版社, 1983
[8] 余金生, 李裕伟. 地质因子分析. 地质出版社, 1985
[9] 于志钧, 赵旭东. 石油数学地质. 石油出版社, 1986
[10] 赵旭东等. 中国数学地质. 地质出版社, 1986
[11] 廖昭懋, 杨文礼. 概率论与数理统计. 北京师范大学出版社, 1988
[12] 王仁铎, 胡光道. 线性地质统计学. 地质出版社, 1988
[13] 卢忠政, 毛鹤琴(主编). 运筹学. 中国建筑工业出版社, 1988
[14] 陆明德, 田时芸. 石油天然气数学地质. 中国地质大学出版社, 1991
[15] 董连科. 分形理论及其应用. 辽宁科学技术出版社, 1991
[16] 林鸿溢, 李映雪. 分形论. 北京理工大学出版社, 1992
[17] 赵旭东. 石油数学地质概论. 石油工业出版社, 1992
[18] 邓昭镜. 超微粒与分形. 西南师范大学出版社, 1993
[19] 张泽厚. 统计学——前沿. 发展. 建设. 中国统计出版社, 1993
[20] 李克文. 利用岩性薄片研究砂岩储层孔隙的分形结构. 石油勘探与开发, 第 20 卷, 第 3 期, p22 ~ 25, 1993
[21] 李后强, 汪富泉. 分形理论及其在分子科学中的应用. 科学出版社, 1993
[22] 沈平平. 利用毛管压力曲线计算孔隙结构的分形维数. 石油勘探与开发, 第 21 卷, 第 2 期, p34 ~ 37, 1994
[23] 毛宁波, 信荃麟, 刘泽容. 应用分形和模式识别技术进行油气预测——以金湖凹陷西 1 断块为例. 江汉石油学院学报, 第 16 卷(增刊), p67 ~ 71. 1994
[24] 徐国祥. 统计预测和决策. 复旦大学出版社, 1994
[25] 丑纪范等. 非线性动力学. 气象出版社, 1994
[26] 金强, 曾怡. 储集性砂岩粒度组成的分形结构. 石油大学学报(自然科学版), 第 19 卷, 第 3 期, p12 ~ 15, 1995
[27] 张济忠. 分形. 清华大学出版社, 1995
[28] 仪垂祥. 非线性科学及其在地学中的应用. 气象出版社, 1995
[29] 施泽进, 罗蛰潭, 彭大钧, 贺振华. 四川盆地油气田空间分布的分形特征. 地球科学——中国地质大学学报, 第 20 卷, 第 1 期, p68 ~ 72, 1995
[30] 刘波. 天地人巨系统观. 安徽教育出版社, 1995
[31] 申维, 赵鹏大. 分形统计模型的理论研究及其在地质学中的应用, 地质科学, 33 卷, 2 期, p234 ~ 243, 1998

[32] 康永尚.沉积盆地的非线性演化特点及其研究意义,中国科协第三届青年学术年会论文集——资源环境与可持续发展技术(高德利主编).中国科学技术出版社,p457 ~ 459,1998.

[33] 康永尚.油气藏规模分布模型和地质模型的关系初探.地球学报,第 20 卷(增刊),p544 ~ 550,1999

[34] 时保宏,康永尚.用二项概率分布模型分析多井钻探风险.西北油气勘探,第 1 期,2002

[35] 时保宏,韩宗文,康永尚.新区勘探的多井钻探计划地质风险分析.西安石油大学学报,19(2),p21 ~ 24,2004

[36] Falconer K.,曾文曲,刘世耀译.分形几何.东北工学院出版社,1991

[37] Turcott,D.L.,陈颙等译.分形与混沌——在地质学和地球物理学中的应用.地震出版社,1993

[38] J.C.Davis.Statistics and data analysis in geology.2nd edition,Wiley.1982

[39] A. Journal.Non-parametric estimation of spatial distributions. Mathematical Geology, V.15, No. 3, p445 ~ 468,1983

[40] M.M.David and B.A.Peter.Scale-dependent fractal dimensions of topographic surfaces:An Empirical Investigation, with Applications in Geomorphology and Computer Mapping.Mathematical Geology, Vol 16, No.7,p671 ~ 683,1984

[41] E.J.Dudewice and S.N.Mishra. Morden mathematical statistics,Wiley.1988

[42] M.F.Barnsley.Fractals everywhere.Orlando,Academic Press,1988

[43] E.D. Isaaks and R.M.Srivastava.Applied geostatistics.Oxford University Press,1989

[44] T.Dewers and P. Ortoleva.The role of geochemical self-organization in the migration and trapping of hydrocarbons.Applied Geochemistry,V.3,p287 ~ 316,1988

[45] B.B.Mandelbrot.What is fractal, and why fractal should matter to petroleum geologist.AAPG,V. 75, No.5,1991

[46] K. Falconer. Fractal Geometry——Mathematical foundations and applications. New York, Wiley,1990

[47] P.R.Rose.Chance of success and its use in petroleum exploration.In:The business of petroleum exploration(Edited by R.Steinmetz).AAPG treatise of petroleum geology,1992

[48] P.J.Ortoleva. Basin compartmentation: definition and mechanisms, In: Basin compartments and seals (Edited by P.J.Ortoleva). AAPG Memoir 61, p39 ~ 51,1994

[49] J.M.Yarus and R.L.Chambers (eds).Stochastic modeling and geostatistics—principles, methods and case studies. AAPG Computer Applications in Geology No.3,1994

[50] Z.Al-Shaieb,J.O.Puckette,A.A.Abdalla and P.B.Ely.Megacompartment complex in the Anadarko Basin: a completely sealed overpressured phenomenon, In: Basin Compartments and Seals, Ortoleva,P.J.(ed). AAPG Memoir 61, p.55 ~ 68,1994

[51] R.A.Crovelli and C.C.Barton.Fractals and Pareto - distribution applied to petroleum accumulation-size distributions, In: Fractals in Petroleum Geology and Earth Processes, C.C.Barton and P.R.La Pointe(eds). Plenum Press, New York and London,p59 ~ 72,1995

[52] P.R.La Pointe.Estimation of undiscovered hydrocarbon potential through fractal geometry, In: Fractals in Petroleum Geology and Earth Processes, C.C.Barton and P.R.La Pointe(eds).Ple-

num Press ,New York and London, p35 ~ 57,1995

[53] J.L.Jensen, L.W,Lake, P.W.M. Corbett and D.J.Goggin .Statistics for petroleum engineers and geoscientists.Prentice Hall PTR,1997

[54] C.C. Barton and R.L.P. Paul.Fractals in the earth sciences. Plenum Press, New York, p141 ~ 178,1995

[55] A. R. H. Swan and M.Sandilands. Introduction to geologist data analysis. Blackwell Science, 1995